Feature-Based Induction

Steven A. Sloman

University of Michigan

Submitted for publication.

Steven Sloman
Dept. of Psychology
330 Packard Rd.
Ann Arbor, MI, 48104-2994

e-mail: sloman@psych.stanford.edu

Abstract

A connectionist model of argument strength is proposed that applies to categorical arguments involving natural categories and predicates about which subjects have few prior beliefs. An example is *robins have sesamoid bones, therefore falcons have sesamoid bones*. The model is based on the hypotheses that argument strength (i) increases with the overlap between features of the combined premise categories and features of the conclusion category; and (ii) decreases with the amount of prior knowledge about the conclusion category. 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.

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 the premise categories are similar to members of the lowest level category that includes both the premise and the conclusion categories." 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, in press, and Smith, Lopez, & Osherson, in press).

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., Loves 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) as well as three other phenomena and also provides good fits to argument-strength ratings.

Osherson et al.'s model is *category-based* in that it assumes 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. (in press), 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 apply to 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 process under consideration, although surely we can come to use it given sufficient training or, possibly, sufficient thought.

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) is discussed. In the course of the discussion, three new phenomena are introduced, one of which is consistent with the feature-based model but not the similarity-coverage model. 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 extentions of the feature-based one.

The similarity-coverage model of argument strength

My presentation of the similarity-coverage 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, $k_1...k_n$, and another category member, g, all of which reside at the same hierarchical level. $SIM_S(k_1...k_n;g)$ is defined as $MAX\{SIM_S(k_1;g),..., SIM_S(k_n;g)\}$. In other words, the similarity between $k_1...k_n$ 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 $k_1...k_n$, $SIM_S(k_1...k_n;G)$ is defined as the average of

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

"...$SIM_S(k_1...k_n;G)$ is the average similarity that S perceives between $k_1...k_n$ and members of G at the level of $k_1...k_n$", (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:

$$P_1...P_n/C,$$

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 category 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 each of $P_1,...,P_n$ and C is a subset of K. 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

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

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 non-blank predicates can be found in this paper's concluding section. Otherwise, my discussion is limited to blank predicates, whose identity is irrelevant and will therefore 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., in press, and Osherson et al., in press). 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

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

The process by which the feature-based model arrives at a judgment of the strength of this argument is diagrammed in Figure 1. The state of the network before presentation of the argument is illustrated in Figure 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 (Figure 1b). To encode the premise of Argument (b) for instance, the units encoding features of Robins 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 (Figure 1c). In the example, this would be accomplished by observing the activation value of unit X (denoted a_X) upon presentation of the features of Falcons.

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, equation (1) below, defines how connections are established. The activation rule, equation (3) below, determines the value of unit X. Because connection strengths depend on the values of features in the premises, a_X increases with the extent of shared features between premise and conclusion categories and decreases with the extent of features in the conclusion category alone, as illustrated in Figure 1c.

Throughout the paper, vectors will be represented by capital letters (e.g., vector B) and their scalar elements will be represented by corresponding lower-case 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(Robins), . . ., f_n(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 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. Because this will always be implicit in what follows, no such indices will be used.

Feature values used for modeling can be derived by having subjects list or rate features or attributes for each category. 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 $A(Y) = [a_1(Y),...,a_n(Y)] = F(Y)$; the current stimulus A(Y) is set equal to the stored representation F(Y). The activation of the output unit, a_X, represents the willingess to confirm the 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, $W = [w_1,...,w_n]$. 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 $W(P_1,...,P_j)$. 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 $a_X(P_j/P_1,...,P_{j-1})$. When testing a conclusion, a_X depends on encoded premises P_1 to P_j and conclusion C, and I will write $a_X(C/P_1,...,P_j)$. 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. Rumelhart, Hinton, & McClelland, 1986) 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,

$$w_i(P_0,P) = w_i(P_0) + \alpha[1 - a_X(P/P_0)]a_i(P). \qquad (1)$$

To keep each weight between 0 and 1, we set

$$\alpha = 1 - w_i(P_0). \qquad (2)$$

This removes the one free parameter we otherwise would have had. Notice that weights never decrease as long as a_X does not exceed one.

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 connection 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 I which is either another premise or the conclusion. The activation rule is

$$a_X(I/P_1,\ldots,P_j) = W(P_1,\ldots,P_j) \bullet A(I) \,/\, |A(I)|^2, \qquad (3)$$

where • means the dot or inner product, defined as $X \bullet Y = \Sigma_{i=1\ldots n} x_i y_i$. The vertical bars represent the length of a vector. By "magnitude", I refer to the denominator of (3), or length squared: $|X|^2 = \Sigma_{i=1\ldots n} x_i^2$. 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 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.

<u>Psychological interpretation of activation rule</u>

Equation (3) states that the activation value of unit X increases with the dot product of the current input and the 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 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 the property is held constant.

I tested the magnitude prediction by asking subjects to choose which of the following two arguments seemed more convincing to them:

Fact: Bobcats have sesamoid bones.
________________________________ (c)
Conclusion: Moles have sesamoid bones.

Fact: Bobcats have sesamoid bones.
________________________________ (d)
Conclusion: Rats have sesamoid bones.

According to feature ratings collected by Osherson et al. (in press),[1] more properties were rated as having a higher strength of association to Rats than to Moles, $|F(\text{Rats})|^2 > |F(\text{Moles})|^2$. This is consistent with the claim that people's representations of Rats are, on average, richer than those of Moles. Subjects were asked to consider each argument on its own merit, ignoring all information other than the facts presented. As predicted by the feature-based model, people preferred the argument with the lower-magnitude conclusion category: 26 out of 40 University of Michigan undergraduates chose Argument (c) over Argument (d) ($p < .05$). This result obtained even though the similarity between categories in the first argument was not significantly greater than the similarity of

---

[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.

categories in the second argument. Following the argument-choice task, similarity ratings were collected from the same 40 students on a 7-point scale where 7 was most similar. The mean similarity rating of Bobcats and Moles was 3.10, which did not differ significantly from the mean similarity rating of Bobcats and Rats (2.95; $t(39) = 1.03$, n.s.).

The magnitude prediction also finds support in people's tendency to generalize more from a member of an out-group to other out-group members than from a member of an in-group to other in-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.

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.

<u>Single premise arguments</u>

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 equation (3), the value of a_x as the first premise is encoded is

$$\begin{aligned} a_x(P_1) &= W() \bullet A(P_1) / |A(P_1)|^2 \\ &= \Sigma\, w_i()\, a_i(P_1) / \Sigma\, a_i(P_1)^2 \\ &= 0, \end{aligned}$$

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

$$\begin{aligned} w_i(P_1) &= w_i() + [1 - w_i()][1 - a_x(P_1)]a_i(P_1) \\ &= 0 + (1 - 0)(1 - 0)a_i(P_1) \\ &= a_i(P_1) = f_i(P_1). \end{aligned} \tag{4}$$

Therefore, the weight vector after encoding a single premise is identical to the vector representing the category in the premise, $W(P_1) = F(P_1)$. Furthermore, the strength of an argument consisting of a single premise and a conclusion is, using first (3) and then (4):

$$a_x(C/P_1) = W(P_1) \bullet F(C) \,/\, |F(C)|^2$$
$$= F(P_1) \bullet F(C) \,/\, |F(C)|^2. \qquad (5)$$

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 ($\mathbf{Proj}_F W$). $\mathbf{Proj}_F W$ is a vector that can be thought as the "F-component" of W (see Figure 2a). It is defined as

$$\mathbf{Proj}_F W = \frac{W \bullet F}{|F|^2} F.$$

Given an argument with premises $P_1,\ldots,P_j$, j applications of the encoding rule will produce weight vector $W(P_1,\ldots,P_j)$. Given conclusion C and conclusion category vector F(C), equation (3) tells us that the strength of the argument is

$$a_x(C/P_1,\ldots,P_j) = W(P_1,\ldots,P_j) \bullet F(C) \,/\, |F(C)|^2.$$

Substituting into the definition of the projection,

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

Because the weights are a (non-linearly derived) representation of the premises, the projection represents the featural components of the premises that are also in the conclusion. It includes those features of the conclusion that overlap with the features of any of the premises as they are encoded in the weights. By calculating the lengths of the two vectors (one on each side of the equation) and re-arranging, we see that

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

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 on 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 Figure 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 on the conclusion category vector.

Also depicted in Figure 2b is the posited difference between arguments (c) and (d) above. Although the projection of F(Bobcats) onto F(Moles) may be equal to the projection of F(Bobcats) onto F(Rats), argument (c) will be stronger than argument (d) according to the model because argument (c)'s conclusion category, F(Rats), is longer than that of argument (d), F(Moles). One interpretation of Figure 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

In order to make contact with the phenomena to be described, a theoretical measure of the similarity between two category representations is required. Perhaps the simplest such measure would be the dot product between the corresponding 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):

$$\mathrm{sim}(D,E) = \frac{F(D) \bullet F(E)}{|F(D)|^2\ |F(E)|^2}. \qquad (6)$$

The virtues of this measure are that it captures most of the properties normally attributed to similarity judgments and, with respect to the feature-based model, it is simple mathematically. Comparing equations (6) and (5), we see that the only difference between sim(P,C) and $a_X(C/P)$ is the presence of an additional term, $|F(P)|^2$, in the denominator of the similarity measure. As we will see, this will facilitate many of the demonstrations 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. It differs from the contrast model in assuming similarity to be a symmetric relation. I am not claiming that equation (6) is a better model of similarity than the contrast model or other reasonable models. The intention of equation (6) is only to describe a relation which, when it holds, has certain consequences for the feature-based model. The feature-based model does not require a model of similarity; similarity does not enter into any of its computations. A model of similarity is needed here only to link the feature-based model of argument strength to phenomena described in terms of similarity.

Since F(D) and F(E) may depend on the context in which they are presented, 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 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. 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 counter-examples 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. & 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 a contrasting pair 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 includes the others, for example

Mice have X
Antelopes have X
____
Aardvarks have X.

Here, Mammal properly includes Mice, 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.		B.
Robins have X	is stronger than	Robins have X
Sparrows have X		Ostriches have X .

Since Robins are more similar to Sparrows than to Ostriches,
sim(Robins, Sparrows) > sim(Robins, Ostriches)
or, from (6),

$$\frac{F(\text{Robins}) \bullet F(\text{Sparrows})}{|F(\text{Robins})|^2\ |F(\text{Sparrows})|^2} > \frac{F(\text{Robins}) \bullet F(\text{Ostriches})}{|F(\text{Robins})|^2\ |F(\text{Ostriches})|^2}$$

which implies

$$F(\text{Robins}) \bullet F(\text{Sparrows}) / |F(\text{Sparrows})|^2 > F(\text{Robins}) \bullet F(\text{Ostriches}) / |F(\text{Ostriches})|^2. \quad (7)$$

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

$$a_X(\text{Sparrows/Robins}) = F(\text{Robins}) \bullet F(\text{Sparrows}) / |F(\text{Sparrows})|^2.$$

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

$$a_X(\text{Ostriches/Robins}) = F(\text{Robins}) \bullet F(\text{Ostriches}) / |F(\text{Ostriches})|^2.$$

Equation (7) tells us that the former value is larger than the latter value so that Argument A is stronger than Argument B.

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

A.		B.
Robins have X		Robins have X
Eagles have X	is stronger than	Eagles have X
Sparrows have X		Ostriches have X,

in which both Robins and Eagles are more similar to Sparrows than to Ostriches. 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 after encoding the premises. The more similar the conclusion category is to the premise categories, the greater will be this dot product. Let R stand for Robins, E for Eagles, S for Sparrows, and O for Ostriches. In the appendix, I show that we can write the strength of Argument A as

$$a_X(S/R,E) = \frac{F(R)\bullet F(S) + [1 - a_X(E/R)]\,[F(E)\bullet F(S) - \sum f_i(R)f_i(E)f_i(S)]}{|F(S)|^2}$$

$$= \frac{F(R)\bullet F(S)}{|F(S)|^2} + [1 - a_X(E/R)]\frac{F(E)\bullet F(S)}{|F(S)|^2} - [1 - a_X(E/R)]\frac{\sum f_i(R)f_i(E)f_i(S)}{|F(S)|^2}$$

and the strength of Argument B as

$$a_X(O/R,E) = \frac{F(R)\bullet F(O) + [1 - a_X(E/R)]\,[F(E)\bullet F(O) - \sum f_i(R)f_i(E)f_i(O)]}{|F(O)|^2}$$

$$= \frac{F(R)\bullet F(O)}{|F(O)|^2} + [1 - a_X(E/R)]\frac{F(E)\bullet F(O)}{|F(O)|^2} - [1 - a_X(E/R)]\frac{\sum f_i(R)f_i(E)f_i(O)}{|F(O)|^2}.$$

Argument strength increases with the featural overlap of both premise categories to the conclusion category as projected in their dot products (e.g., $F(R)\bullet F(S)$ and $F(E)\bullet F(S)$ 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., $\sum f_i(R)f_i(E)f_i(S)$). 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.

As we did in the single premise case, we can show that the two dot product terms in Argument A are greater than the corresponding terms in Argument B because Sparrows are more similar to both Robins and Eagles than are Ostriches. 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, as long as the magnitude of the Sparrow vector is not much greater than that of the Ostrich vector. Moreover, unlike the first dot product term, the effect of the three-way product is diminished by the coefficient preceding it $[1 - a_X(E/R)]$. Therefore, 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 of premises. The same encoding rule, equation (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 provide a good match to 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.

<u>ii. & iii. Premise diversity (General and specific).</u>

The less similar are premises to themselves, 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. The account of the specific case follows from that of the general one. The feature-based model attributes the diversity phenomenon to feature coverage. More diverse premises cover the space of features better than more similar premises because their features are not redundant, and are therefore more likely to be encoded and more likely to overlap with features of the conclusion category. Specifically, (i) weight changes are greater during encoding of a dissimilar premise than a similar premise (dissimilar premises have more surprise value); and (ii) 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. 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 implies

$$a_X(\text{Rhinos/Hippos}) > a_X(\text{Hamsters/Hippos}). \qquad (8)$$

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

$$a_X(M/Hi,Ha) = \frac{F(Hi)\bullet F(M) + [1 - a_X(Ha/Hi)]\,[F(Ha)\bullet F(M) - \sum f_i(Hi)f_i(Ha)f_i(M)]}{|F(M)|^2}$$

and the strength of Argument B as

$$a_X(M/Hi,R) = \frac{F(Hi)\bullet F(M) + [1 - a_X(R/Hi)]\,[F(R)\bullet F(M) - \sum f_i(Hi)f_i(R)f_i(M)]}{|F(M)|^2} .$$

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

$$[1 - a_X(Ha/Hi)]\,[F(Ha)\bullet F(M) - \sum f_i(Hi)f_i(Ha)f_i(M)] > [1 - a_X(R/Hi)]\,[F(R)\bullet F(M) - \sum f_i(Hi)f_i(R)f_i(M)]$$

which I now show is likely to hold. Inequality (8) tells us that $[1 - a_X(Ha/Hi)] > [1 - a_X(R/Hi)]$. Because most of us comparing these arguments do not know substantially more about Rhinos than about Hamsters, and because Rhinos are not obviously more typical Mammals than are Hamsters, we can expect that for most of us $F(R)\bullet F(M)$ is not substantially greater than $F(Ha)\bullet F(M)$. The only remaining difference between the two sides of the inequality lies in the three-way product terms $\sum f_i(Hi)f_i(Ha)f_i(M)$ and $\sum f_i(Hi)f_i(R)f_i(M)$. 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, $\sum f_i(Hi)f_i(R)f_i(M)$ is likely to be greater than $\sum f_i(Hi)f_i(Ha)f_i(M)$. So a bigger quantity is being subtracted from a smaller quantity in Argument B relative to Argument A and Argument A is therefore likely to be stronger. More precisely, Argument A will be stronger than Argument B for every subject for whom these conditions hold.

<u>Feature exclusion (new phenomenon).</u>

The feature-based model predicts a boundary condition on the diversity phenomenon. A premise category that has no overlap with the conclusion category should have no effect on argument strength even if it leads to a more diverse set of premises. Consider the argument

Fact: German Shepherds have sesamoid bones.
Fact: Blue whales have sesamoid bones.
______________________________ (e)
Conclusion: Moles have sesamoid bones.

According to Osherson et al.'s (in press) feature ratings, there is almost no overlap between the features of Blue whales and those of Moles (the dot product was only 1.6). Let GS be German Shepherds, BW be Blue whales, and M be Moles. The derivation in the appendix tells us that the argument strength of (e) is

$$a_X(M/GS,BW) = \frac{F(GS) \bullet F(M) + [1 - a_X(BW/GS)][F(BW) \bullet F(M) - \sum f_i(GS) f_i(BW) f_i(M)]}{|F(M)|^2}.$$

If $F(BW) \bullet F(M)$ is negligible, then so is $\sum f_i(GS) f_i(BW) f_i(M)$ since it is always positive but less than $F(BW) \bullet F(M)$ which implies that

$$\begin{aligned} a_X(M/GS,BW) &\approx F(GS) \bullet F(M) / |F(M)|^2 \\ &= a_X(M/GS). \end{aligned}$$

In other words, Blue whales should not contribute strength to the argument, even though German Shepherds and Blue whales comprise a diverse set. To test this analysis, I asked 60 University of Michigan undergraduates to choose the stronger of argument (e) and the following argument which has a less diverse set of premises:

Fact: German shepherds have sesamoid bones.
Fact: Giraffes have sesamoid bones.
________________________________ (f)
Conclusion: Moles have sesamoid bones.

The first argument indeed has a more diverse set of premises than the second; the similarity between German Shepherds and Blue Whales was rated significantly lower than the similarity between German Shepherds and Giraffes on a 7-point scale (2.25 and 3.25 respectively, $t(59) = 5.67$; $p < .001$). Nevertheless, the feature-based model predicts Argument (f) will be stronger than Argument (e) because the featural overlap between Giraffes and Moles is not negligible (dot product of 3.1). As predicted, 48 out of 60 students chose (f).

The similarity-coverage model can explain this result by assuming that, even though the premises in (e) had greater diversity, they had less coverage. If Blue whales have few similar neighbors, they may not add much to the coverage of the lowest-level category that includes both premises and conclusion.[2] But to derive the prediction from the similarity-coverage model, we need to first define the lowest-level category both with and without Blue whales

---

[2]Supporters of the similarity-coverage model cannot appeal to greater similarity between the premises and conclusion of Argument (e) relative to Argument (f) 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 40 subjects rated Giraffes more similar to Moles than German Shepherds to Moles. With regard to the next demonstration, only 7 of 39 subjects rated Beavers more similar to Chihuahuas than Foxes to Chihuahuas. When these subjects are excluded from the analysis, only a single subject preferred argument (g).

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, all we need to know is that the feature overlap between the premise and conclusion categories is negligible.

I replicated the above demonstration using a different pair of arguments:

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

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

The diversity between the premises of Argument (g) is greater than that in Argument (h); 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 out of 59 students chose (h). This was predicted by the feature-based model 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.

<u>iv. Premise typicality (General).</u>

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

A.		B.
Robins have X		Penguins have X
_____________	is stronger than	_____________ ,
Birds have X		Birds have X

where Robins are more typical of the category Birds than are Penguins. By virtue of being more typical, the similarity of Robins to Birds is greater than the similarity of Penguins to Birds. Therefore, from (6),

$$\text{sim(Robins,Birds)} = \frac{F(\text{Robins}) \bullet F(\text{Birds})}{|F(\text{Robins})|^2\ |F(\text{Birds})|^2}$$

$$> \frac{F(\text{Penguins}) \bullet F(\text{Birds})}{|F(\text{Penguins})|^2\ |F(\text{Birds})|^2} = \text{sim(Penguins,Birds)}.$$

Tversky (1977) proposed that the features of more typical items are more salient or prominent. I will make the weaker assumption that the representations of typical categories are, in general, at least as rich as the representations of atypical categories, so that $|F(Robins)|^2 \geq |F(Penguins)|^2$, which implies, by substituting into the above inequality,

$$F(Robins) \bullet F(Birds) / |F(Birds)|^2 > F(Penguins) \bullet F(Birds)./ |F(Birds)|^2. \qquad (9)$$

To derive the strength of Argument A, we use equation (5):

$$a_X(Birds/Robins) = F(Robins) \bullet F(Birds) / |F(Birds)|^2.$$

The strength of Argument B is

$$a_X(Birds/Penguins) = F(Penguins) \bullet F(Birds) / |F(Birds)|^2.$$

The inequality in (9) dictates that $a_X(Birds/Robins) > a_X(Birds/Penguins)$ or that Argument A is stronger than Argument B.

Because the explanation for this phenomenon relies on similarity as well as richness, rather than typicality per se, it makes a prediction distinct from that of the similarity-coverage model. It suggests that if a pair of categories D and E can be found 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.

v. & 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

A.		B.
Sparrows have X		Sparrows have X
Eagles have X		Eagles have X
Hawks have X	is stronger than	
______________		______________
Birds have X		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(Birds/Hawks,Sparrows,Eagles)

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

and the strength of Argument B is

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

Recall that weights are never decreased (assuming $a_X \leq 1$). Adding a premise will always increase weights connected to features of the new premise 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

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

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

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

<u>vii. Inclusion fallacy (One specific and one general argument).</u>

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:

A.		B.
Robins have X		Robins have X
___________	is stronger than	___________ .
Birds have X		Ostriches have X

The feature-based model predicts this phenomenon whenever the similarity of the premise category to the general category (Birds) is greater than its similarity to the specific category (Ostriches). The account is much like the account of phenomenon i., similarity. Using definition (6),

$$\text{sim(Robins,Birds)} = \frac{F(\text{Robins}) \cdot F(\text{Birds})}{|F(\text{Robins})|^2 \ \ |F(\text{Birds})|^2}$$

$$> \frac{F(\text{Robins}) \cdot F(\text{Ostriches})}{|F(\text{Robins})|^2 \, |F(\text{Ostriches})|^2} = \text{sim(Robins,Ostriches)},$$

which implies that

$$F(\text{Robins}) \cdot F(\text{Birds}) \,/\, |F(\text{Birds})|^2 > F(\text{Robins}) \cdot F(\text{Ostriches}) \,/\, |F(\text{Ostriches})|^2. \qquad (10)$$

The strength of Argument A, derived as above, is

$$a_X(\text{Birds/Robins}) = F(\text{Robins}) \cdot F(\text{Birds}) \,/\, |F(\text{Birds})|^2,$$

and the strength of Argument B is

$$a_X(\text{Ostriches/Robins}) = F(\text{Robins}) \cdot F(\text{Ostriches}) \,/\, |F(\text{Ostriches})|^2.$$

Since condition (10) holds, a_X(Birds/Robins) > a_X(Ostriches/Robins) or Argument A is stronger than Argument B.

To observe the inclusion fallacy, the model requires that the premise category selects or overlaps with more features of the conclusion category in the general argument than in the specific argument relative to the overall number of features in each conclusion category. This can be seen most clearly by the following restatement of (10), which assumes that F(Robins) • F(Ostriches) > 0:

$$\frac{F(\text{Robins}) \cdot F(\text{Birds})}{F(\text{Robins}) \cdot F(\text{Ostriches})} > \frac{F(\text{Birds}) \cdot F(\text{Birds})}{F(\text{Ostriches}) \cdot F(\text{Ostriches})}.$$

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 since Birds is more specific than Animals and Sparrows and Eagles are Birds which are in turn Animals

A.		B.
Sparrows have X		Sparrows have X
Eagles have X	is stronger than	Eagles have X
________________		______________ .
Birds have X		Animals have X

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, and the ordering of similarity judgments is consistent with equation (6), the feature-based model expects the conclusion specificity phenomenon to reverse. 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". Gelman (1988) has offered an alternative explanation for this phenomenon 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:

A.		B.
Tigers have X	is stronger than	Buffaloes have X
Buffaloes have X		Tigers have X .

The phenomenon has been attributed to differences in typicality (Osherson et al., 1990; Rips, 1975). 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. The phenomenon follows if Tigers are assumed more typical than Buffaloes of some common category, such as Mammals.

The feature-based model attributes the phenomenon to differential richness of category representations (which is likely correlated with typicality). According to Osherson et al.'s (in press) feature ratings, more properties are strongly associated to Tigers than to Buffaloes; the vector representation of Tigers has greater magnitude than that of Buffaloes, $|F(\text{Tigers})|^2 > |F(\text{Buffaloes})|^2$. From (5), we write the strength of Argument A as

$$a_X(\text{Buffaloes/Tigers}) = F(\text{Tigers}) \bullet F(\text{Buffaloes}) \,/\, |F(\text{Buffaloes})|^2,$$

and the strength of Argument B as

$$a_X(\text{Tigers/Buffaloes}) = F(\text{Buffaloes}) \bullet F(\text{Tigers}) \,/\, |F(\text{Tigers})|^2.$$

Because the magnitude of the vector representing Tigers is larger than that representing Buffaloes,

$$a_X(\text{Buffaloes/Tigers}) > a_X(\text{Tigers/Buffaloes})$$

because their numerators are identical. Therefore, Argument A is stronger than Argument B.

x. Premise-conclusion identity.

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

Pelicans have X
_____________ is perfectly strong.
Pelicans have X

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

$$
\begin{aligned}
a_X(\text{Pelicans/Pelicans}) &= F(\text{Pelicans}) \bullet F(\text{Pelicans}) \,/\, |F(\text{Pelicans})|^2 \\
&= \sum f_i(\text{Pelicans}) f_i(\text{Pelicans}) \,/\, |F(\text{Pelicans})|^2 \\
&= |F(\text{Pelicans})|^2 \,/\, |F(\text{Pelicans})|^2 \\
&= 1.
\end{aligned}
$$

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

Pelicans have X
Flamingoes have X
____
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:

Animals have X
______________ is perfectly strong. (i)
Mammals have X

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 (i) is

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

The numerator will tend to increase as the similarity between the premise and conclusion categories increases, and therefore so will the 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.

Premise-conclusion 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:

1 A.
Animals use norepinephrine as a neurotransmitter.
____
Mammals use norepinephrine as a neurotransmitter.

1 B.
Animals use norepinephrine as a neurotransmitter.
____
Reptiles use norepinephrine as a neurotransmitter.

Forty-four out 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

2 A.
Birds have an ulnar artery.
___
Robins have an ulnar artery.

2 B.
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:

3 A.
Furniture cannot be imported into Zimbabwe.
___
Tables cannot be imported into Zimbabwe.

3 B.
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 out of 7.

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 premise-conclusion 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 (SE = .28), significantly greater than the mean rating for Argument 1B (Animals therefore Reptiles) of 5.88 (SE = .38), $t(59) = 5.07$; $p < .001$. The same pattern held for the second pair of arguments (Birds therefore Robins versus Birds therefore Penguins). Mean convincingness ratings were 9.45 (SE = .14) and 6.73 (SE = .37), respectively, $t(59) = 7.35$; $p < .001$. The third pair also showed the phenomenon. The mean for 3A (Furniture therefore Tables) of 9.32 (SE = .19) was significantly higher than the mean for 3B (Furniture therefore Bookshelves) of 8.53 (SE = .24), $t(59) = 2.86$; $p < .01$. The argument with the more similar categories was judged significantly more convincing in all three cases. Similarity judgments replicated the patterns reported above.

These demonstrations support the feature-based model which predicts the strength of arguments satisfying premise-conclusion inclusion to be proportional to similarity and refute the category-based model which cannot explain these data without auxilliary assumptions. One such auxilliary assumption[3] 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. The support provided by the premise-conclusion inclusion similarity phenomenon for the feature-based model should be qualified however. The model predicts the phenomenon only to the extent that equation (6) is consistent with the ordering of similarity judgments.

<u>xii. & xiii. Nonmonotonicity (General and specific).</u>

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	is stronger than	Peacocks have X
		Rabbits have X
___________		___________
Birds have X		Birds have X.

Similarly,

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

---

[3] Suggested by Ed Smith.

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 possibility is that these nonmonotonicities are the result of factors that neither model is intended to capture. The phenomena could result from a kind of pragmatic violation. One type of pragmatic violation is discussed by Osherson et al. (1990): The unrelated category, like Orangutans, appears irrelevant to the conclusion. They provide some evidence against this possibility. Another, more plausible, kind of pragmatic violation that would explain the results is that the unrelated premise category is so distinctive that subjects infer that the sort of thing which has property X must come from a strange category that they cannot identify and this inhibits them from generalizing to new categories.

Another potential explanation, 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. Subjects may only be willing to encode relations amongst those features that are consistent with all premise categories. Consider a three-step mechanism to encode premises. First, a set of units receive input from all premises so that the system has access to the union of premise features. Second, a recurrent network sends activation back and forth amongst these units to maximize their internal consistency or "harmony" (Smolensky, 1986). This kind of competitive process can be found in a variety of connectionist models (e.g., Anderson, Silverstein, Ritz, & Jones,1977). Third, only those units whose activations survive the competition are connected to the predicate unit. In this way, only features maximally consistent with all premise categories would be encoded so that nonmonotonicities would be expected whenever the additional premise category has features inconsistent with many features of the other premise and conclusion categories.

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. I used the feature lists constructed by Osherson et al. (in press), who obtained ratings of "the relative strength of association" between a set of properties and a set of mammals. These ratings are not ideal for testing the feature-based model. The appropriate kind of judgment would instead ask 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. Some association strength can be found between Sheep and *has tail* or *is weak* in Osherson et al.'s rating data because these properties are associated with the category even though they may not be salient aspects of subjects' representations. For this reason, judgments of relative strength of association seem to encourage 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 equations (1), (2), and (3). Judgments of argument strength were obtained from data published in Osherson et al. (1990) and Smith et al. (in press). 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., in press, 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.

Correlations were calculated between the model's predictions using feature ratings, calculated for each cutoff value, and subjects' probability estimates. Because cutoff values had only two decimal-places precision, the cutoff parameter space was essentially discrete, which allowed for the evaluation of correlations at all possible cutoff points. 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.

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. (in press) 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.

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., in press, 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. (in press) report a multiple correlation of 0.96 between the two estimated terms of their model and probability judgments.

The third data set consisted of specific arguments of, again, two premises each (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.[4]

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 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. (in press, 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, the data were not used to estimate any parameter intrinsic to the model, neither a learning nor an activation parameter. To model specific arguments, the similarity-coverage model requires a

---

[4]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.

parameter specifying the relative weight of the similarity and coverage terms. Second, 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. Third, for reasons given above, the feature ratings were not optimal for testing the feature-based model. Fourth, lacking subjects' direct judgments, feature ratings had to be estimated for two categories (Dog and Mammal). Finally, 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 one category given confirmation of the property in one or more categories. 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, premise-conclusion inclusion similarity, does. The model also predicted another new phenomenon, feature exclusion. Some support was offered for a third new phenomenon, category richness, in which I showed that an argument with a conclusion category whose representation had a lower magnitude was preferred over one with a higher magnitude. 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 determing 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. In Osherson et al.'s (1990) data, the asymmetry phenomenon proved the hardest to demonstrate. On the other hand, only four of the phenomena depended on differences in 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 amongst 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 some feature-based model were consistent with a wide variety of data under many different conditions, some 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 amongst categories is not necessarily derivable from a feature-based model which appeals to relations amongst 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.

Third, the models make differing predictions about the effect of premise order. Because the feature-based model uses the delta rule of equation (1), premises are encoded only to the extent that they are surprising. Small differences can therefore sometimes be expected between an argument of the form P_1,P_2/C and one of the form P_2,P_1/C. For instance, if the set of features of the category in P_2 are properly included in the set of features of the category in P_1, then the latter argument is predicted to be 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.

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 micro-level processing mechanisms.

The problem of non-blank predicates

The most important advantage of the feature-based view, and of the model proposed here in particular, is the obvious direction it suggests for generalizing to arguments involving non-blank predicates. Non-blank predicates would be modeled as consisting of features, some of which have pre-existing 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 of predicates that they do not already know.

A second implication of this type of feature-based model for non-blank predicates is that arguments will tend to be judged strong given a strong prior belief in their conclusion. 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

Onions are white.

____

Therefore, George Bush is white.

may be judged weak for strategic reasons. For example, subjects might decide that the features of onions do not overlap with those of George Bush. These issues will have to be clarified before we can expect a feature-based model for non-blank 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.

References

Anderson, J. A., Silverstein, J. W., Ritz, S. A., & Jones, R. S. (1977). Distinctive features, categorical perception, and probability learning: Some applications of a neural model. *Psychological Review, 84*, 413-451.

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 Rumelhart, D. E., McClelland, J. L., & the PDP Research Group (Eds.) *Parallel Distributed Processing, 1*, The MIT Press, Cambridge, MA.

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. (in press). Default probability. *Cognitive Science*.

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.

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

Smith, E. E., Lopez, A., & Osherson, D. (in press). Category membership, similarity, and naive induction. To appear in A. Healy, S. Kosslyn, & R. Shiffrin (Eds.) *Essays in honor of W. K. Estes.*

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.

Smolensky, P. (1986). Information processing in dynamical systems: Foundations of harmony theory. In Rumelhart, D. E., McClelland, J. L., & the PDP Research Group (Eds.) *Parallel Distributed Processing, 1*, The MIT Press, Cambridge, MA.

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.

Rumelhart, D. E., Hinton, G. E., & McClelland, J. L. (1986). A general framework for parallel distributed processing. In Rumelhart, D. E., McClelland, J. L., & the PDP Research Group (Eds.) *Parallel Distributed Processing, 1*, The MIT Press, Cambridge, MA.

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

Appendix

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

D have X
E have X
____
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 (5),

$$a_X(E/D) = F(D) \bullet F(E) \,/\, |F(E)|^2.$$

Using equations (1) and (2), each weight i after having encoded both premise categories is

$$w_i(D,E) = w_i(D) + [1 - w_i(D)]\,[1 - a_X(E/D)]\,f_i(E)$$

$$= f_i(D) + [1 - f_i(D)]\,[1 - a_X(E/D)]\,f_i(E).$$

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

$$a_X(C/D,E) = W(D,E) \bullet F(C) \,/\, |F(C)|^2$$

$$= \frac{\sum f_i(D)f_i(C) + \sum \{[1 - f_i(D)]\,[1 - a_X(E/D)]f_i(E)f_i(C)\}}{|F(C)|^2}$$

$$= \frac{F(D)\bullet F(C) + [1 - a_X(E/D)]\,[F(E)\bullet F(C) - \sum f_i(D)f_i(E)f_i(C)]}{|F(C)|^2}$$

Author Notes

This research was conducted while the author held a post-doctoral fellowship from the Natural Sciences and Engineering Research Council of Canada. I thank Edward Wisniewski, Douglas Medin, and especially Edward E. Smith for helpful comments and discussion.

Correspondence concerning this article should be addressed to Steven Sloman, Department of Psychology, Univeristy of Michigan, 330 Packard Road, Ann Arbor, Michigan 48104-2994.

Figure 1. Illustration of feature-based model for the argument Robins have X, therefore Ravens have X.

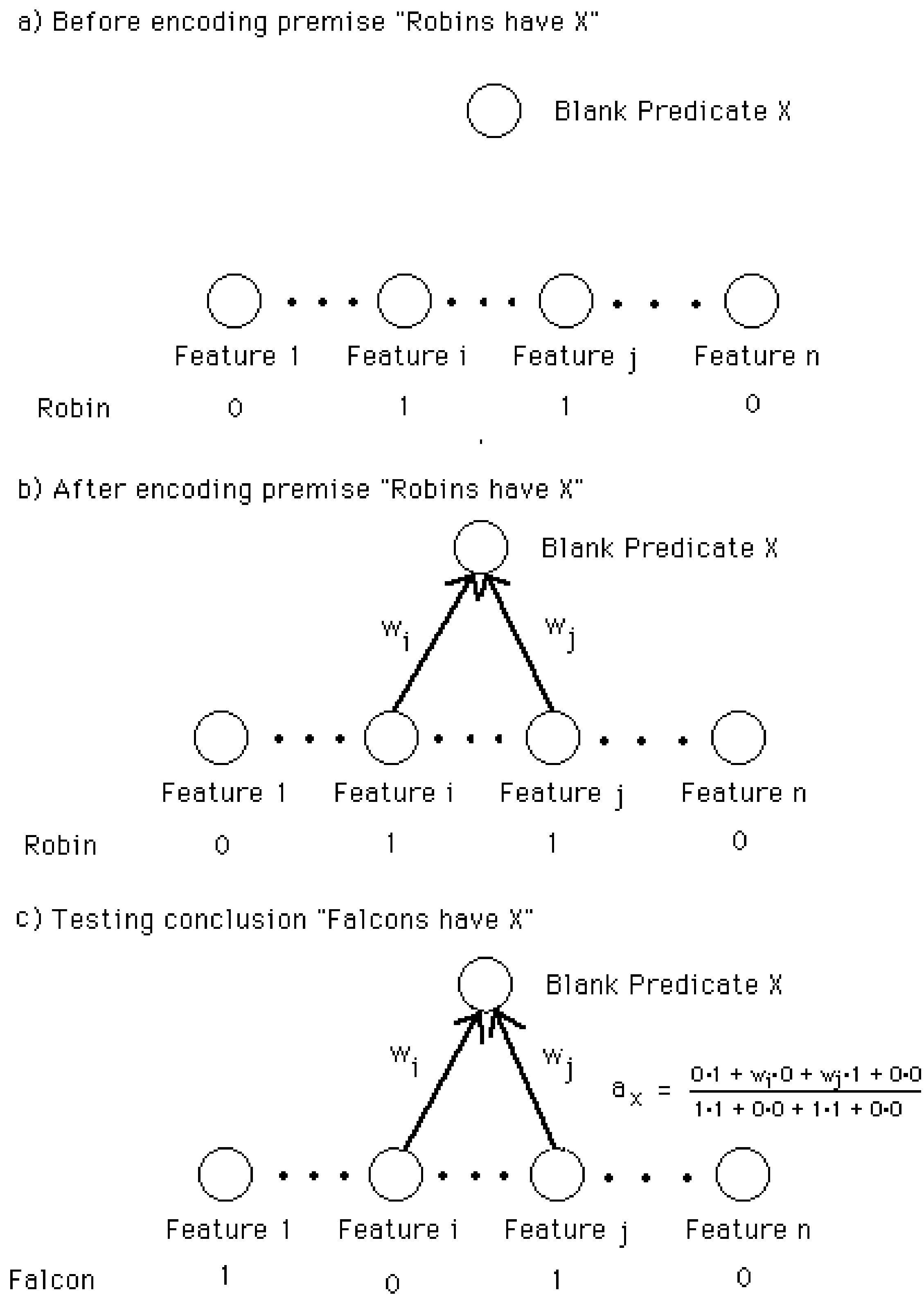
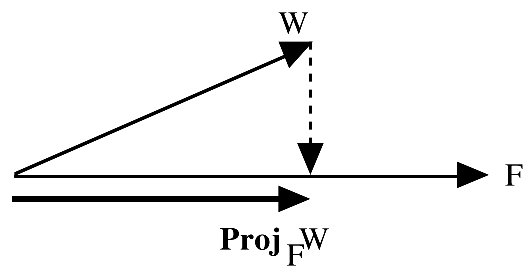


Figure 2. Geometric interpretation of the feature-based model.

a) General case



$$a_x = \frac{\text{Length}(\mathbf{Proj}_F W)}{\text{Length}(F)}$$

b) Illustration of single-premise arguments