COGNITIVE SCIENCE **15,** 251–269 (1991)

Default Probability

DANIEL N. OSHERSON
I.D.I.A.P.

JOSHUA STERN
ORMOND WILKIE
Massachusetts Institute of Technology

MICHAEL STOB
Calvin College

EDWARD E. SMITH
University of Michigan

A probability may be called "default" if it is neither derived from preestablished probabilities nor based on considerations of frequency or symmetry. Default probabilities presumably arise through reasoning based on causality and similarity. This article advances a model of default probability based on a featural approach to similarity. The accuracy of the model is assessed by comparing its predictions to the probabilities provided by undergraduates asked to reason about mammals.

1. INTRODUCTION

One of the most fundamental cognitive acts is the attribution of a probability $p(S)$ to a statement S, for example, the attribution of .7 probability to the claim that the economy will weaken next year. We may distinguish four ways in which people produce such attributions.

1. *Relative Frequency.* Having observed a sample of m individuals, n of which have a certain property, it is common to ascribe probability n/m to the statement that another individual drawn from the same population will also possess the given property. A large literature testifies to the fact that people often rely in this way on relative frequency as a guide to probability, even in the absence of random sampling (see Estes, 1976 for a review).

Research support was provided by a Siemens Corporation grant to Osherson, by National Science Foundation Grant Nos. 8609201 and 8705444 to Osherson and Smith, respectively, and by the Office of Naval Research under contract No. N00014-87-K-0401 to Osherson.

Correspondence and requests for reprints should be sent to Edward E. Smith, Human Performance Center, The University of Michigan, 330 Packard Road, Ann Arbor, MI 48104.

2. *Principles of Symmetry.* Given one face of a cubical and homogeneous die, it is natural to assign probability 1/6 to the statement that this face will turn up after a vigorous role. Such an intuition is based on the symmetry of the die and on some version of the doctrine of "insufficient reason." (For justification and extension of this doctrine see Jaynes, 1979. Myers & Osherson, in press, offer discussion from a psychological point of view.)
3. *Derivation from Preestablished Probabilities.* People often attempt to deduce desired probabilities from probabilities antecedently attributed to statements, for example, using Bayes' theorem. Such mental deductions presuppose an inference procedure, perhaps implicitly held, and perhaps deviant from the viewpoint of classical probability theory. A considerable body of psychological research has been devoted to characterizing the inference procedures that underlie derivations of this kind. (See Baron, 1988 for a review.)
4. *Default Reasoning.* If a desired probability cannot be ascertained using the foregoing methods, it is necessary to rely on reasoning schemas of a nonprobabilistic kind, involving causal inference and similarity. Collins and Michalski (1989) examined a variety of schemas of this kind, but did not connect them to probability estimation per se.

This article addresses default reasoning about probability. In particular, we consider judgments about the (conditional) probability of statements concerning mammals, given the truth of other statements. A similarity-based model of such reasoning is advanced and evaluated against judgments elicited from undergraduates. Although numerous schemes have been advanced for reasoning by similarity (see Vosniadou & Ortony, 1989), there appear to be no proposals for converting similarity into specific probabilities. A successful method of this kind would be a contribution not only to psychology, but also to artificial intelligence inasmuch as it would help to isolate analogical processes in automated reasoning, focusing them solely on default probabilities. Inference can then be carried out within the framework of classical probability theory. (See Pearl, 1988, Section 1.4, for the advantages of this strategy in automated inference.)

Our goal is limited to showing the feasibility of converting similarity into probability, rather than advancing the definitive similarity model. Consequently, we shall attempt to demonstrate the predictive power of a simple model of this kind, treating more complex alternatives cursorily. Although our model is based primarily on similarity, we do not deny the importance of causal schemas and other nonsimilarity mechanisms in probabilistic reasoning. (See Collins & Michalski, 1989 for discussion of many such principles.) Indeed, by assessing the strengths and weaknesses of similarity approaches to default probability, the role of nonsimilarity mechanisms may be expected to emerge more clearly.

The underlying idea of our model can be conveyed as follows. Suppose that objects $o_1 \ldots o_n$ each have property P, and that none of $o_1' \ldots o_m'$ have P. Then, in the absence of other information, the probability that some new object o has P is assumed to vary directly with the similarity of o to $o_1 \ldots o_n$ and inversely with the similarity of o to $o_1' \ldots o_m'$. Several principles are needed in order to make this idea precise. For the case in which all the objects are at the same hierarchical level, we need principles that determine (a) the similarity between pairs of objects; and (b) the amalgamation of multiple, pairwise similarities into an overall judgment. For the case in which objects are at different hierarchical levels, we need additional principles that determine (c) the decomposition of higher-level objects into lower-level ones.

Principles relevant to (b) and (c) will be derived from the theory of category-based induction advanced in Osherson, Smith, Wilkie, Lopez, & Shafir (1990). Knowledge of the latter theory is not presupposed here, however, because the needed principles will be introduced later. With regard to (a), we rely on a feature-based conception of similarity. Given mammals m_1, m_2 with feature sets M_1, M_2, the similarity of m_1 to m_2 is taken to be:

$$(1) \quad sim(m_1, m_2) = \frac{M_1 \cap M_2}{M_1 \cup M_2}$$

This model has a long history in psychology and biology (see Gregson, 1975, Section 2.5). Its accuracy in this context is documented in a separate experiment reported later.

We now overview the empirical studies used to test the model. All the studies center on 48 mammals, chosen for familiarity and diversity; they are listed in Table 1. Eighty-five properties were selected to represent common knowledge about the 48 mammals. Abbreviations for the properties are listed in Table 2 (p. 254), and sample properties are given in unabbreviated form in Table 3 (p. 254). Subjects always worked with unabbreviated properties; the abbreviations are for expositional ease. With the exception of animal noises (bleating, roaring, etc., essentially unique to each animal), no other property was listed by more than a single subject from a group of 10 MIT

TABLE 1
Mammals

antelope	deer	horse	persian cat	spider monkey
bat	dalmatian	humpback whale	pig	squirrel
beaver	fox	leopard	polar bear	tiger
blue whale	german shepard	lion	rabbit	walrus
bobcat	giant panda	killer whale	raccoon	weasel
buffalo	giraffe	mole	rat	rhinoceros
chihuahua	gorilla	moose	seal	wolf
chimpanzee	grizzly bear	mouse	sheep	zebra
collie	hamster	otter	siamese cat	
elephant	hippopotamus	ox	skunk	

TABLE 2
Abbreviated Properties

black	white	blue	brown	gray	orange
red	yellow	patches	spots	stripes	furry
hairless	toughskin	big	small	bulbous	lean
flippers	hands	hooves	pads	paws	longneck
longleg	tail	chewteeth	meatteeth	buckteeth	strainteeth
horns	claws	tusks	smelly	flys	hops
swims	tunnels	walks	fast	slow	strong
weak	muscular	bipedal	quadrupedal	active	inactive
nocturnal	hibernates	agile	eats fish	eats meat	eats plankton
eats vegetation	eats insects	forager	grazer	hunter	scavenger
skimmer	stalker	newworld	oldworld	arctic	coastal
desert	bush	plains	forest	fields	jungle
mountains	ocean	ground	water	tree	cave
fierce	timid	smart	group	solitary	nestpot
domesticated					

TABLE 3
Sample, Unabbreviated Properties

black:	the color black in its visual appearance
bulbous:	having a roundish or bulky body shape
longleg:	having a long leg
chewteeth:	having molars that are good for chewing
vegetation:	commonly eats vegetation in its natural habitat
newworld:	living in the New World (North and South America)
agility:	having a high degree of physical coordination
swims:	swimming as a means of locomotion
ocean:	living in the ocean
nestspot:	keeping their young in a designated, enclosed area

students asked to supply properties of mammals. Moreover, none of the 85 properties were judged to be inappropriate by more than 1 student in the same group. These pilot studies, along with the coherence of the results reported later, suggest that the 85 properties capture common knowledge about familiar mammals.

Three rating tasks were performed in this study, each employing a separate group of subjects. The first task measured the strength of association between each of the 48 mammals and each of the 85 properties. The second task obtained similarity ratings between pairs of mammals. The third task focussed on probability judgment. The property-rating task is described in Section 2. Its purpose was to build a database of mammal facts from which similarity between mammals could be calculated. The ability to predict similarity on this basis is tested in the similarity task, described in Section 3. The probability task is described in Section 4. Its purpose was to assess different methods for generating default probabilities by comparing generated prob-

abilities with the actual judgments of our subjects. Ideally, every subject would have rated every mammal on every property, and also completed the similarity and probability tasks. In practice, it was decided that subjects should work for no more than an hour, participating in just one of the tasks. This procedure minimizes fatigue and the risk of contaminating judgments in one task by recollection of another. On the other hand, the accuracy of our analyses are thereby limited by the effects of between-subject variability, as will be pointed out later.

The methods and results of the three tasks are now described. All subjects were MIT undergraduate volunteers, recruited through advertisements and paid for their participation.

2. PROPERTY-RATING TASK

2.1 Method

Subjects first reviewed the list of 48 mammals and 85 properties (unabbreviated). It was explained that a nonnegative number was to be assigned to each mammal–property pair; the number should reflect "the relative strength of association between the property and the mammal." No upper bound was imposed on these ratings. Subjects were also told to expect that many of the properties would be negligibly associated with many of the mammals. A rating of 0 was to be used for these cases.

Each subject worked for 1 hour, evaluating 10–15 randomly chosen mammals on all 85 properties (faster subjects evaluated more mammals). For each mammal evaluated, all the properties were rated for that mammal before the next mammal was introduced. Properties were rated in the order given in Table 2 (each row read from left to right). A computer terminal was used to present properties and record data. Subjects worked individually at their own speed and had the opportunity to review and revise their prior ratings at any time. Twenty-nine subjects participated in the property-rating task. Random sampling of the mammals was constrained so that each mammal was evaluated by 8 or 9 subjects.

2.2 Results

So that averages would not be biased by those subjects using large numbers, every subject's ratings were individually normalized by a linear transformation to range from a lowest score of 0 to a highest score of 1. For each mammal, the normalized scores of the 8 or 9 subjects rating it were averaged. The result is a 45×85 matrix whose i, j-cell approximates the degree to which property j is associated with mammal i in the minds of MIT undergraduates. Henceforth this matrix will be denoted by M. The ith row of M corresponds to the ith mammal of the 48 used in the study; this mammal will be denoted by m_i.

The following statistics provide some information about the variability of the ratings for the different mammals. The overall association to m_i is defined as

$$\Sigma_{j=1}^{85} M(i, j).$$

The average overall association to the 48 mammals is 17.76 ($SD = 2.63$). The number of nontrivial associations to m_i is defined to be the number of $j \leq 85$ such that $M(i, j) \geq .1$. The average number of nontrivial associations to the 48 mammals is 41.9 ($SD = 5.64$).

3. SIMILARITY TASK

3.1 Method

To test the psychological reality of similarity model (1) as well as the inter-subject stability of our Mammal × Property matrix, 30 subjects were asked to rate the similarity of pairs of mammals drawn from the initial stock of 48.[1] None of the 30 subjects had participated in the property-rating task. The following instructions were employed:

> This experiment concerns your judgment about the biological similarity of different mammals. The similarity of two mammals depends on how alike they are in physiology, anatomy, diet, behavior, habitat, appearance, etc. For each pair of mammals that is presented, you will assign a value between 0 and 100 (decimals allowed) that reflects the similarity that you perceive between the mammals mentioned in the pair. Numbers closer to 100 should reflect greater similarity, numbers closer to 0 should reflect lesser similarity.

For each subject 40 pairs of mammals were individually randomly selected with the sole constraints that (a) no identity pairs (e.g., zebra–zebra) be included; and (b) no two pairs of the form x–y and y–x be included. A given subject's 40 pairs were sequentially presented for rating on a computer terminal in randomized order. The mammals of a pair appeared on the same line, the choice of left-most mammal being determined randomly. Subjects worked at their own speed and could review and revise earlier ratings at any time. The procedure typically lasted 30 minutes.

3.2 Results

We define the following function *sim* from pairs of mammals to [0, 1]. Given mammals m_i, m_k.

$$(2) \quad sim(m_i, m_k) = \frac{\Sigma_{j=1}^{85} \mathit{minimum}\{M(i, j), M(k, j)\}}{\Sigma_{j=1}^{85} \mathit{maximum}\{M(i, j), M(k, j)\}}$$

[1] Two additional subjects were excluded from the experiment because they responded incorrectly to at least 3 of 10 elementary questions about mammals administered in a preexperimental interview.

The *sim* function (2) reduces to the *sim* function (1) of Section 1 if the association $M(i, j)$ of property j to mammal i is conceived as consisting of "microfeatures" that sum to $M(i, j)$; the greater the number of such microfeatures, the greater the level of association. The minimum of $\{M(i, j), M(k, j)\}$ may then be conceived as the intersection of two sets of microfeatures, and the maximum as their union. The intersection represents the commonality of m_i. m_k, whereas the union is the sum of commonality and distinctiveness. [Distinctiveness is computed by

$$\Sigma_{j=1}^{85} \mid M(i, j) - M(k, j) \mid .]$$

Definition (2) has three features that render it more appropriate to this study than Tversky's (1977) well-known contrast model of similarity, which places commonality and distinctiveness in linear combination. First, definition (2) ensures that similarities, like probabilities, are numbers in [0, 1]. In comparison, the contrast model allows similarities to be any number, positive or negative. Our attempt to derive probability from similarity will be facilitated by the restricted range of the similarity function (2). Second, definition (2) implies that for every mammal m_i, $sim(m_i, m_i) = 1$, which corresponds to the maximum informativeness of m_i in inferences about m_i. In comparison, the contrast model allows $sim(m_i, m_i)$ to be any positive real number, and $sim(m_i, m_i) \neq sim(m_j, m_j)$ is possible for distinct mammals m_i, m_j. There seems to be no fact about inference that corresponds to this variability in self-similarity. Finally, no free parameters appear in definition (2), whereas three are required for the contrast model. The absence of parameters simplifies the evaluation of models in what follows.

For each of the 30 subjects we computed the Pearson correlation between (a) the similarity values assigned by that subject to the 40 pairs of mammals he or she evaluated, and (b) the values of *sim* for those same pairs, computed from (2). Note that *sim* values do not depend on any data from the similarity subjects, because only the matrix M enters their calculation and M was constructed from the data of the property-rating task. As a consequence, between-subject variability in opinions about mammal features can be expected to lower the correlation between observed similarity values and predicted *sim* values. Nevertheless, the average of these 30 correlations is .64 ($p < .001$, $SD = .123$). We interpret this result as supporting the psychological reality of the Mammal × Property matrix M as well as definition (2) of similarity.[2]

4. PROBABILITY TASK

The probability task consisted of a categorization procedure followed by a judgment procedure. The purpose of the first procedure was to identify the

[2] Because each subject received an individually randomized set of pairs for rating, no analysis using pooled data is possible.

superordinate categories that the subject recognizes among mammals. These superordinates figured in the probability questions to which the subject responded in the second procedure. Thirty new subjects completed the probability task.[3] Before describing the two procedures we discuss the nature of the probability questions used.

4.1 Probability Questions Used

4.1.1 General Form of the Questions. In the judgment procedure subjects evaluated conditional probabilities like those appearing in the following questions.

(3) (a) What is the probability that horses require biotin for hemoglobin synthesis assuming that giraffes do?
(b) What is the probability that all canines use norepinephrine as a regulator of sexual drive assuming that wolves do and felines do not?
(c) What is the probability that all mammals can regulate their feeding cycle in conditions of constant illumination assuming that bears can?

The statement "All canines use norepinephrine as a regulator of sexual drive" will be called the *conclusion* of question (3 b), whereas the succeeding statements about wolves and felines will be called *premises,* and similarly for other questions. As illustrated in (3 b), some questions included negative premises. Conclusions were always affirmative.

The premises and conclusion of a given question always invoke the same predicate and have one of the following logical forms: (a) all members of category X have property P, or (b) all members of category X do not have property P. The predicates figuring in the questions, for example, "requires biotin for hemoglobin synthesis," may be termed *blank* inasmuch as subjects are unlikely to attach prior probabilities to conclusions involving such properties. The use of blank predicates thus allows all relevant background information to appear explicitly in the premises of a probability question. This study is limited to blank predicates; extension to nonblank predicates is briefly discussed in Section 6.2. Probability questions will henceforth be abbreviated by (a) omitting their predicates, (b) writing premises above conclusion with a separating line, and (c) indicating premise polarity by + or −. Thus, (3 b) is abbreviated to:

$$\begin{array}{l} +\ \textit{wolves} \\ \underline{-\ \textit{felines}} \\ \textit{canines} \end{array}$$

[3] One additional subject was excluded from the experiment because he failed the preexperimental test described in Footnote 1. Another additional subject was dropped because of highly bizarre superodinates (his data were not analyzed).

Question (3 b) illustrates the presence of superordinate categories like *feline* and *canine* among premises and conclusions. These superordinates do not figure among our list of 48 mammals but rather include subsets of them. Pilot studies revealed that superordinates recognized by M.I.T. undergraduates are variable in both name and membership. Consequently, with the exception of *mammal* (assumed common to everyone), the superordinates figuring in a given subject's questions were drawn exclusively from the set established in that subject's categorization procedure.

A vast number of probability questions may be generated from the 48 mammals plus associated superordinates. Each subject responded to an individually randomly selected subset of questions that met certain criteria. One criterion excluded defective questions; other criteria included questions of suitable type. The next two subsections set forth these criteria.

4.1.2 Exclusion of Defective Questions. Three kinds of defective probability questions are now defined. (The definitions are relative to the superordinate categories established by a particular subject.) A question is *contradictory* if its premises cannot all be true. Suppose, for example, that both the superordinates *feline* and *man-eating* contain *lion*. Then question (4) is contradictory.

(4) $+$*felines*
$-$*man-eaters*
rhinos

A question is *redundant* if one of its premise categories includes another. For example, (5) is redundant if *canine* is the standard category.

(5) $+$*canines*
$+$*German shepards*
rabbits

Similarly, $\{+$*lion*, $+$*lion*$\}$ is a redundant premise set.

A question is *trivial* if its premises logically imply its conclusion, or the negation of its conclusion. For example, the following are trivial (assuming that *canine* is the standard category).

(6) $-$*collies*
$+$*foxes* / *canines* $+$*canines* / *wolves*

We also consider trivial any question whose conclusion is implied by its premises under the assumption that our 48 instances exhaust the category *mammal*. By this criterion (7) is trivial, if the union of *predator* and *prey* includes all 48 mammals.

(7) $+$*predators*
$+$*prey*
mammals

All probability questions posed to subjects were noncontradictory, nonredundant, and nontrivial.

4.1.3 Inclusion of Suitable Types. A premise or conclusion may be called "specific" if its category is one of the 48 mammals of Table 1; it is "superordinate" if its category is defined by the subject as including at least 2 but not all of the 48 mammals; and it is "general" if its category is *mammal*. We distinguish four types of premises: either specific or superodinate, and either affirmative or negative (general premises are excluded by nontriviality). We distinguish three types of conclusion, either specific, superordinate, or general (all conclusions are affirmative). Two probability questions are said to be of the *same type* just in case (a) their conclusions are of the same type; and (b) the number of premises of each type are equal across the two questions. For example, the following pairs of questions are of the same type:

+ bobcat	*− feline*
− canine	*+ rat*
seal	*skunk*
+ elephant	*− lion*
− sheep	*+ hamster*
+ primate	*+ canine*
canine	*feline*
+ beaver	*+ collie*
mammal	*mammal*

A counting argument shows that there are exactly 47 types of questions meeting the following conditions:

(8) (a) the question has 1, 2, or 3 premises;
(b) it has at least 1 positive premise; and
(c) it has at most one negative premise.

Any question of one of these 47 types—provided that it is neither contradictory, redundant, nor trivial—was potentially available for use in the probability task. We now describe the categorization and probability procedures that constituted the task.

4.2 Categorization Procedure

Subjects first read the following instructions:

> This part of the experiment concerns your judgment about how to distribute mammals into natural categories. Your task will be to create biologically meaningful groups, and then for each group to indicate which of the 48 mammals belongs to it. It is permitted to leave a mammal uncategorized if there are no other mammals in the list with which it forms a biologically natural group.

> Groups can be of any size, and it is permissible to have overlap of members. For each group, you will need to devise a short, descriptive label.

Categorization was carried out on a computer terminal. Subjects devised category names and indicated which mammals among the 48 were included in it. Review and revision of previous choices of category name and membership was possible at any time. The superordinate name "mammal" was not allowed. The categorization procedure lasted roughly 30 minutes.

4.3 Judgment Procedure

Subjects first read the following instructions:

> This part of the experiment concerns your judgment about the probability that a category of mammals possesses a given, biological property. The properties in question might involve any biologically meaningful aspect of mammals, including their physiology, anatomy, diet, behavior, habitat, appearance, etc. Examples of these properties are the following:
>
> - requires biotin for hemoglobin synthesis;
> - has sesamoid bones;
> - can regulate their feeding cycle in conditions of constant illumination;
> - blood salinity declines from infancy to maturity;
> - uses norepinephrine as a regulator of sexual drive.
>
> Imagine that a biological property like one of these has recently come under investigation. You know nothing about the property except that it is biological in character, and called *"P"* for short. You will be asked to judge the probability that one kind of mammal has property *P*, assuming it to be known that other kinds of mammals do—or do not—have *P*.

Forty-seven probability questions were then randomly generated for each subject. The superordinates appearing in the questions were drawn from the list established by the same subject in the preceding categorization procedure. Each question exemplified a distinct type from the set of 47 types satisfying (8). No question was either contradictory, redundant, or trivial. Within these constraints, the mammals and superordinates appearing in a given question were chosen randomly for each subject individually. For multiple-premise questions, the order of premises was determined randomly. The order in which a given subject's 47 questions were presented for evaluation was also determined randomly.

The judgment procedure was carried out on a computer terminal. Questions appeared in the form exemplified by (9).

(9) Given that:
(1) Rats have the property *P*,
(2) no canines have the property *P*,
(3) felines have the property *P*,
what is the probability (0–100%) that all primates have the property *P*?

After responding to the 47 questions, subjects reviewed their answers and could revise any of them.

5. RESULTS OF THE PROBABILITY TASK

5.1 Global Statistics

The average number of superordinates generated by the 30 subjects in the categorization procedure was 11.4 ($SD = 3.49$), with a minimum of 5 and a maximum of 20. Over all 30 subjects, the average number of mammals included in a given superordinate was 5.8 ($SD = 4.66$). The average probability assigned by a given subject in the judgment procedure ranges from .220 to .620. Over all subjects, the mean of these averages is .477 ($SD = .118$).

5.2 Assessing Default Reasoning Models: General Remarks

We now consider several models for predicting the probabilities assigned by an individual subject. All the models are assessed as follows. One question is selected from the 47 evaluated by a given subject. It is the probability assigned to this "target" question that must be predicted. The prediction is generated by whatever computation is prescribed by the model at issue. This computation may use as input no more than: (a) the subject's answers to the 46 remaining questions; (b) information about membership in the subject's superordinate categories (as established in the categorization procedure for that subject); and (c) the Mammal × Property matrix M established in the property-rating task. The absolute difference between predicted and assigned probabilities for the target is determined. A new target question is then selected and the remaining 46 questions (including the old target question) are used to generate a prediction about the new target. This procedure is repeated for all 47 possible target questions. The performance of the model for the given subject is measured by the average, absolute deviation over all 47 questions between predicted and assigned probabilities. This average is called the *discrepancy* for the chosen subject. We seek a model that minimizes the average discrepancy across all 30 subjects.

5.3 An Actuarial Model

In order to establish baseline performance for comparison with other models, an actuarial model for generating default probabilities was assessed. To predict the probability assigned by the subject to the target question we used the average probability assigned by that subject to the remaining 46 questions. The average discrepancy for this model across all 30 subjects is .191 ($SD = .047$).[4]

---

[4] Because each of the 47 questions that a subject answered was of unique type (in the sense of Section 4.1.3), it is not possible to predict a target question by averaging over the subset of remaining questions of the same type. Such an averaging scheme might be the best actuarial model in a context where multiple questions of the same type were evaluated by a single subject.

5.4 A Similarity Model

Our similarity model was briefly discussed in Section 1, and may be described as follows. Let a target question Q be given. We first determine the similarity of Q's positive premises to its conclusion as well as the similarity of Q's negative premises to its conclusion. The probability assigned to Q is taken to be a linear combination of these latter two similarities. The coefficients of the linear combination are derived by regression over the remaining 46 questions. We now describe this procedure precisely.

The *sim* function of definition (2) applies to pairs of mammals. To implement our model of probability judgment we must extend *sim* to a function *SIM* defined on pairs X, Y of subsets of mammals. Intuitively, $SIM(X, Y)$ measures the extent to which X "covers" Y, specifically, the extent to which every member of Y is near to some member of X. *SIM* is defined as follows.

(10) Let X, Y be subsets of mammals, and let y be a particular mammal.
  (a) $SIM(X,y) = maximum\{sim(x,y) \mid x \in X\}$;
  (b) $SIM(X, Y) = mean\{SIM(X,y) \mid y \in Y\}$.

Thus, $SIM(X,y)$ is the maximum similarity of a member of X to y, and may be termed "the similarity of X to y." $SIM(X, Y)$ is the average similarity of X to members of Y. [Note that $SIM(X, Y)$ need not equal $SIM(Y,X)$.]

As a means of predicting our subjects' assigned probabilities, *SIM* has some noteworthy properties. Three of these are now discussed. Let Q be a probability question whose premises are positive and specific, and whose conclusion is superordinate. Let $X = x_1 \ldots x_n$ be the mammals figuring in the premises, and let $Y = y_1 \ldots y_m$ be the mammals included in the conclusion category.

1. $SIM(X, Y)$ is monotone in n, as easily seen. Likewise, the probability that subjects actually assign to Q usually does not decline with expansion of the premise set X. Exceptions to this generalization are documented in Osherson et al. (1990) under the term "nonmonotonicity." The exceptions are rare enough, however, to warrant the monotonicity of *SIM*.
2. *SIM* is not monotone in m; The *mean* operator in (10 b) allows $SIM(X, Y)$ either to increase or decrease as Y is expanded. Normatively, we expect the probability of Q to decline monotonically with increasing m. But subjects often violate this principle when faced with questions like the following, judging the first to be more likely than the second.

$$\frac{+\,mouse}{mammal} \qquad \frac{+\,mouse}{hippo}$$

This pattern of judgment is documented in Osherson et al. (1990) under the term "inclusion fallacy." In contrast to nonmonotonicity with respect to premises, the inclusion fallacy is prevalent in ordinary reasoning (see Shafir, Smith, & Osherson, in press). The use of *mean* in (10 b)

is a simple mechanism for representing this feature of naive judgment. For example, $SIM(\{mouse\}, mammal) > SIM(\{mouse\}, \{hippo\})$, because almost all mammals resemble mice more than hippos do.[5]

3. *SIM* conforms to the "diversity effect," namely, the tendency for the probability assigned to Q to rise as the average similarity between members of X declines. This effect is documented in Osherson et al. (1990); it has also been discussed by philosophers of science (e.g., Horwich, 1982). It is easy to see that $SIM(X, Y)$ also tends to rise with the diversity of X (because of the maximum operator in (10 a). For further discussion of these and other properties of *SIM*, see Osherson et al. (1990).

We now describe our similarity model for generating default probabilities for a given subject. The model first associates a *positive similarity factor* and a *negative similarity factor* with each of the subject's 47 questions. For each question Q, these factors are calculated in the following three steps.

Step 1: *Segregate the affirmative and negative premises of Q so as to form two subquestions, denoted Q^+ and Q^-.* For example, if Q is:

(11) $+ox$
$-canine$
$+feline$
seagoing

then Q^+ and Q^- are as follows:

(12) $+ox$
$+feline$ | $-canine$
seagoing | *seagoing*

If Q contains no negative premises, then Q^- is null. (By (8 b), Q has at least one positive premise.)

Step 2: *"Explode" Q^+ and Q^- by replacing superordinate categories with their members (according to the subject's categorization data).* Thus, assuming natural memberships for *feline, canine* and *seagoing*, Q^+ and Q^- from (12) become:

(13) $+ox$

$+bobcat$	$-chihuahua$
$+leopard$	$-collie$
$+lion$	$-dalmatian$
$+persian\ cat$	$-fox$

[5] It is easy to see that use of *minimum* in (10 b) rather than *mean* would block the inclusion fallacy, and hence, is more normatively acceptable. Thus, in modeling the probability judgment of experts instead of undergraduates, a model based on *minimum* would presumably be more descriptively accurate than one based on *mean*.

+siamese cat	*−german shepard*
+tiger	*−wolf*
bluewhale	*bluewhale*
humpback whale	*humpback whale*
killer whale	*killer whale*
seal	*seal*
walrus	*walrus*

Q^- remains null if Q contains no negative premises. If Q's conclusion is general (i.e., contains *mammal*) than all 48 mammals appear in the conclusion set of its exploded arguments. If a premise or conclusion is specific, then explosion does not affect it.

Step 3: *Calculate* $SIM(X^+,C)$ *where* X^+ *is the set of mammals appearing in* Q^+*'s exploded premises, and C is the set of mammals appearing in* Q^+*'s conclusion. If* Q^- *is not null, calculate* $SIM(X^-,C)$ *in the same fashion.*

Q's positive similarity factor is $SIM(X^+,C)$ from Step 3. If Q^- is null, then Q's negative similarity factor is defined to be zero; otherwise, it is $SIM(X^-,C)$. We let Q^{pos} be Q's positive similarity factor and Q^{neg} be Q's negative similarity factor. It is expected, of course, that Q^{pos} vary directly with the judged probability of Q and that Q^{neg} vary inversely.

Finally, given target question Q and remaining questions Q_i ($i \leq 46$), this model assigns a default probability to Q in the following manner. Using standard techniques from the theory of linear regression, real coefficients α, β, γ are found such that

$$(14)\quad \sum_{i=1}^{46} (\alpha Q_i^{pos} + \beta Q_i^{neg} + \gamma - \hat{Q}_i)^2$$

is minimized, where $\hat{Q}_i$ is the probability assigned by the subject to Q_i. The probability predicted for Q is then:

$$(15)\quad \alpha Q^{pos} + \beta Q^{neg} + \gamma$$

The average discrepancy (in the sense of Section 5.2) for this model across all 30 subjects is .152 ($SD = .041$).[6] A t test for related measures shows this performance to be significantly superior to that of the actuarial model of Section 5.3 ($t = 9.09$, $p < .001$). The discrepancy for 29 of the 30 subjects was lower using the similarity model than using the actuarial model.

A related analysis of the similarity model was carried out as follows. For each subject we calculated the multiple correlation between the probability assigned to a given question Q and the values of Q^{pos} and Q^{neg} for that

[6] It is possible for (15) to fall outside the interval [0, 1]. However, this occurs so seldomly that no truncation step was employed to convert negative values to 0 or values greater than 1 to 1. We note as well that in virtually every case, α turned out to be positive and β turned out to be negative, as expected.

question as given in Step 3 above. Thus, for each of the 30 correlations (1 per subject), $N = 47$, which is the number of questions randomly generated for each subject. The average of these 30 correlations is .60 ($SD = .124$). The mean value of the regression coefficient for Q^{pos} was .93; for Q^{neg} it was $-.25$. The discrepancy in absolute value suggests that subjects paid more attention to positive than to negative premises.[7]

As before, it is well to note that the predictions of the similarity model rest heavily on the data of the subjects who rated mammal properties. Specifically, the calculation of Q^{pos} and Q^{neg} for a given question Q depends only on the superordinate categories elicited from the subject in question plus the matrix M used to calculate *sim*. As a consequence, between-subject variability in opinions about mammal properties can only depress the fit of the similarity model to the data of the probability task.

5.5 A Pure Category Model

To gauge the role of similarity per se in the accuracy of the similarity model, we devised a rival model that exploits information about superordinate categories provided by each subject in the probability task, but does not depend on similarity. Thus, the rival model uses only data provided by the subject being modeled, because no recourse is made to *sim* and the matrix M upon which it is based.

Given question Q, let Q^{prem} denote the number of premises in the exploded version of Q^+ minus the number of premises in the exploded version of Q^-. Let Q^{conc} denote the number of conclusions in the exploded version of Q^+ or Q^-. Thus, Q^{prem} measures the weight of evidence in favor of Q's conclusion, whereas Q^{conc} measures the generality of that conclusion. Q^{prem} and Q^{conc} are based entirely on a given subject's category information; similarity does not intervene.

Our category model is the same as the similarity model except that Q^{prem} and Q^{conc} are used in place of Q^{pos} and Q^{neg} respectively.[8] The average discrepancy for this model across all 30 subjects is .179 ($SD = .044$). A t test for related measures shows this performance to be significantly superior to that of the actuarial model ($t = 3.57$, $p < .01$), but significantly worse than that of the similarity model ($t = 7.61$, $p < .001$). The discrepancy for 27 of the 30 subjects was lower using the similarity model than using the pure category model.

5.6 Other Models

We have tried other methods for generating default probabilities, but they either work less well than the similarity model or are more complicated and work no better. The variations that were investigated include the following:

---

[7] Because each subject received an individually randomized sample of 47 arguments, no analysis is possible using pooled data.

[8] Truncation in the sense of Footnote 6 was employed to limit predictions to [0, 1].

1. Substitution of a linear similarity function for definition (2) of *sim;*
2. Differential weighting of common and distinctive properties in calculating *sim;*
3. Enhanced weighting of properties that are shared by several premises in calculating similarity;
4. Replacement of *maximum* by *sum* in (10 a) and replacement of *mean* by either *minimum* or *maximum* in (10 b); and
5. Averaging techniques of various sorts in order to create "prototype vectors" from positive premises, negative premises, and conclusions.

The foregoing variations were also tried in combinations.

6. DISCUSSION

This investigation is preliminary in two respects. First, the experimental procedure limits the accuracy that can be expected of any model of default probability. Second, blank rather than interpretable predicates figured in the probability questions. These topics are now discussed in turn.

6.1 Limits on Accuracy in This Study

6.1.1 Division of Tasks. One set of subjects constructed the Mammal × Property matrix M and a different set of subjects responded to probability questions. This circumstance allows between-subject variability in knowledge about mammals to interfere with predictions of probability judgment. In an application of the similarity model for purposes of generating default probabilities automatically, information about objects and properties would be based on judgments made by the same person whose probabilities are to be predicted.

6.1.2 Limited Number of Mammals. Only 48 mammals figured in this study. As a result, categories like *canine* are likely to include members (e.g., *poodle*) that fall outside the 48 mammals that subjects categorized. The exploded arguments generated in Step 2 of Section 5.4 are, therefore, imperfect representations of questions involving categorical premises and conclusions. A more comprehensive set of instances is likely to arise in a realistic setting. Similarly, realistic databases might code information about property variability (see Nisbett, Krantz, Jepson, & Kunda, 1983; Rips, 1989) and about the typicality of instances; these are potentially useful variables in similarity calculations.

6.1.3 Number of Predictive Variables. Only two variables—namely, positive and negative factors of similarity—appear in the predicting formula (15) of Section 5.4. Other variables might be linearly combined with these two in the hope of increasing predictive accuracy beyond the .152 average

discrepancy achieved. In particular, the theory of category-based induction advanced in Osherson et al. (1990) posits additional variables related to the superordinate categories that subjects recognize among mammals.

It is usually self-defeating, however, to incorporate additional variables into (15) because of the limited number of probability questions evaluated by each subject. For practical reasons, and in order to avoid stereotypical responding, each subject in the judgment procedure responded to only 47 questions. As a consequence, our attempts to add additional predictive variables were foiled by the emergence of ad hoc solutions to the regression equations based on 46 items. Use of these solutions to predict the probability assigned to target questions results in greater average discrepancy than obtained with just two variables, or else yields little improvement at the expense of a more complicated method.

In a more realistic setting, a larger number of judgments would be available, so extrapolation to a new probability can be based on methods incorporating more variables. Nonlinear use of these variables might also be worthwhile.

6.1.4 Minimization of Absolute Differences. The regression analysis used to fix the parameters α, β, and γ in (15) minimizes the squared deviation (14). In contrast, it is more natural to define average discrepancy in terms of absolute deviation, as we have done. The average discrepancy of the similarity method could thus be further reduced by minimizing

$$(16) \qquad \sum_{i=1}^{46} | \alpha Q_i^{pos} + \beta Q_i^{neg} + \gamma - \hat{Q}_i |$$

rather than (14) when fixing α, β, and γ. Minimization of absolute differences is computationally difficult, which is why familiar regression techniques were employed here. However, we have employed numerical techniques to estimate values of α, β, γ that minimize (16). Using these estimates, the average discrepancy of the similarity model is diminished by nearly 10%.

6.2 Nonblank Predicates

Extension of our results to probability questions with meaningful predicates is nontrivial because interactions can arise between a property explicitly ascribed to a given object and other properties it possesses (cf., Murphy & Medin, 1985). An initial approach to meaningful predicates is to limit them to properties already represented explicitly in prestored information about instances (e.g., size, habitat, color, etc. in this study). Attribution of such a property to an instance would change or confirm the value of the property initially represented for that instance. To reflect the greater importance of an explicitly attributed property, its weight in similarity calculations would be increased. A similar technique yielded positive results in a study of typi-

cality and conceptual combination (see Smith, Osherson, Rips, & Keane, 1988). Stern (1991) applies mechanisms of this character to modeling default probability in the context of meaningful predicates.

REFERENCES

Baron, J. (1988). *Thinking and deciding.* Cambridge: Cambridge University Press.

Collins, A., & Michalski, R. (1989). The logic of plausible reasoning: A core theory. *Cognitive Science, 13,* 1-50.

Estes, W.K. (1976). The cognitive side of probability learning. *Psychological Review, 83,* 37-64.

Gregson, R. (1975). *Psychometrics of similarity.* New York: Academic.

Horwich, P. (1982). *Probability and evidence.* Cambridge: Cambridge University Press.

Jaynes, E.T. (1979). Where do we stand on maximum entropy? In R.D. Levine & M. Tribus (Eds.), *The maximum entropy formalism.* Cambridge, MA: MIT Press.

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

Myers, T., & Osherson, D. (in press). On the psychological appeal of the maximum entropy principle.

Nisbett, R., Krantz, D., Jepson, C., & Kunda, Z. (1983). The use of statistical heuristics in everyday inductive reasoning. *Psychological Review, 90,* 339-363.

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

Pearl, J. (1988). *Probabilistic reasoning in intelligence systems.* San Mateo, CA: Morgan-Kaufmann.

Rips, L. (1989). Similarity, typicality, and categorization. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning.* Cambridge: Cambridge University Press.

Shafir, E., Smith, E., & Osherson, D. (in press). Typicality and reasoning fallacies. *Memory and Cognition.*

Smith, E., Osherson, D., Rips, L., & Keane, M. (1988). Combining prototypes: A selective modification model. *Cognitive Science, 12,* 485-527.

Stern, J. (1991). *Default reasoning about probability.* Unpublished manuscript.

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

Vosniadou, S., & Ortony, A. (Eds.). (1989). *Similarity and analogical reasoning.* Cambridge: Cambridge University Press.