DANIEL OSHERSON, EDWARD E. SMITH, TRACY S. MYERS,
ELDAR SHAFIR AND MICHAEL STOB

EXTRAPOLATING HUMAN PROBABILITY JUDGMENT

ABSTRACT. We advance a model of human probability judgment and apply it to the design of an extrapolation algorithm. Such an algorithm examines a person's judgment about the likelihood of various statements and is then able to predict the same person's judgments about new statements. The algorithm is tested against judgments produced by thirty undergraduates asked to assign probabilities to statements about mammals.

Keywords: Probability, judgment, psychology.

The present paper advances a model of human probability judgment and applies it to the design of an extrapolation algorithm. Such an algorithm examines a person's judgment about the likelihood of various statements and is then able to predict the same person's judgments about new statements.

Section 1 describes the kind of extrapolation task for which our model is designed. The model itself is presented in Section 2. Section 3 shows how the model may be used to extrapolate human probability judgment. Concluding remarks occupy Section 4.

1. EXTRAPOLATION TASKS

The extrapolation tasks we consider are built around object-names and predicates. In our experiments, the former refer to mammals, like 'Lions', 'Rabbits', and 'Deer' whereas the latter express biological properties like 'have three layers of lipid tissue surrounding vital organs'. A pair (O, P) consisting of object O and predicate P defines a statement attributing P to O. By an *argument* is meant a statement s associated with a (possibly empty) set $\{s_1 \cdots s_m\}$ of statements such that $s \notin \{s_1 \cdots s_m\}$: s is the argument's *conclusion*, and $s_1 \cdots s_m$ are its *premises*. Arguments may be written in the form $(s \mid s_1 \cdots s_m)$ or vertically as in (1).

Theory and Decision **36**: 103–129, 1994.

TABLE I

Number of m-premise arguments based on five objects and 1 predicate.

m	m-premise arguments
0	5
1	20
2	30
3	20
4	5
total:	80

$$\begin{array}{l} S_1 \\ \vdots \\ \underline{S_m} \\ S \end{array} \qquad (1)$$

We refer to (1) as an 'm-premise' argument. 0-premise arguments are just statements. Note that the premises of an argument are an unordered set. To write an argument, the premises are ordered arbitrarily.

For simplicity in what follows attention is limited to arguments in which only a single predicate appears. Negations and other connectives are also absent. The model can be extended in a natural way to arguments of greater complexity, but this is not done here. For present purposes, the general form of an argument is thus:

$$\begin{array}{l} (O^1, P) \\ \vdots \\ \underline{(O^m, P)} \\ (O, P) \end{array} \qquad (2)$$

Any choice of sets $\mathcal{O}, \mathcal{P}$ of objects and predicates induces a

corresponding set of arguments. For example, if $\mathcal{O}$ has five members and $\mathcal{P}$ has one, then there are five statements and eighty distinct arguments, as described in Table I.

By an *agent* A for $\mathcal{O}$ and $\mathcal{P}$ we mean any system that maps the set of arguments generated by $\mathcal{O}$ and $\mathcal{P}$ into the interval $[0, 1]$. Given such an argument $(s \mid s_1 \cdots s_m)$, $A(s \mid s_1 \cdots s_m)$ may be conceived as the probability that A assigns to s while assuming the truth of $s_1 \cdots s_m$, i.e., the conditional probability for A of s given $s_1 \cdots s_m$. In the case of a 0-premise argument s, $A(s)$ is just the (unconditional) probability that A attributes to s. Since the agent in question may be human, we do not assume that A's probabilities conform to the standard probability calculus (except for being representable by the interval $[0, 1]$).

Now suppose that a set of objects $\mathcal{O}$, a set of predicates $\mathcal{P}$, and a class of agents $\mathcal{A}$ for $\mathcal{O}$ and $\mathcal{P}$ have been specified. The *extrapolation problem* for $\mathcal{O}$, $\mathcal{P}$, and $\mathcal{A}$ is to find an algorithm **alg** that behaves as follows. An agent $A \in \mathcal{A}$ and a finite set of arguments $\alpha_1 \cdots \alpha_k$ based on $\mathcal{O}$, $\mathcal{P}$ are chosen arbitrarily. The pairs $(\alpha_1, A(\alpha_1)) \cdots (\alpha_k, A(\alpha_k))$ are fed to **alg** (each pair can be interpreted as a message of the form: 'To such-and-such argument the chosen agent attributes such-and-such probability'). **alg** Then performs a calculation over the input and enters a state that allows it to compute A to some reasonable approximation in the sense that for all remaining arguments α, $A(\alpha) \approx$ **alg**(α).

Many versions of the extrapolation problem can be distinguished, depending on the number and diversity of the pairs exhibited to **alg,** on the quality of the approximation required, on the likelihood that **alg** delivers the desired approximation, and on the time allowed for **alg** to finish its calculations. It is not necessary in what follows to establish terminology for all these cases. We aim simply to present evidence that the extrapolation problem is solvable – in a general, qualitative sense – when objects and predicates are drawn from a familiar domain like mammals, and when potential agents are limited to college students.

Our extrapolation algorithm rests on a specific model of human probability judgment. The input data are used to fix the parameters of this model so as to simulate the psychology of the unknown agent. Once the parameters have been fixed, the model is used to predict new judgments.

We now consider the model in question.

2. THE GAP MODEL OF PROBABILITY JUDGMENT

2.1. Vectorial Representations

As indicated in the last section our model concerns judgments elicited by arguments like the following.

Lions have three layers of lipid tissue surrounding vital organs.
Rabbits have three layers of lipid tissue surrounding vital organs.
Deer have three layers of lipid tissue surrounding vital organs. (3)

A central assumption of the model is that people's beliefs about objects and predicates in natural domains can be represented by real vectors in an appropriate attribute space. (See [33, 3, 32] for similar assumptions in other models.) Suppose for illustration that a given subject distinguishes three attributes of mammals, namely: *size*, *ferocity*, and *frigidness-of-habitat*. Then the objects of argument (3) might be represented in this mind as shown in Table II.

We assume that predicates can be evaluated along the same dimensions as objects. The predicate of argument (3), for example, might give rise to the last column in Table II. The value '4' in this column represents the size required of mammals to have vital organs surrounded by three layers of lipid tissue, according to the conceptions of our given subject.

It is not assumed that attributes are independent, either conceptually or stochastically. In particular, one attribute might represent the

TABLE II

Hypothetical vectors associated with the three objects and one predicate figuring in argument (3).

	Lions	Rabbits	Deer	3 lipid layers
attributes				
1) *size*	3	1	3	4
2) *ferocity*	6	0	1	0
3) *frigid-habitat*	−3	3	4	5

interaction of two others, for example, as the product of their values. The psychological reality of interactions among attributes has been noted in [15, 16, 17].

Henceforth we use variables like O and P to denote not only grammatical entities like objects and predicates, but also the vectors assumed to be associated with them.

2.2. *Probability of Statements*

For a statement (O, P) to be probable, the values in the vector O should be at least as great as the corresponding values in P. This idea may be quantified with the 'cut-off' operator $\dot{-}$, defined over real numbers by:

$$x \dot{-} y = max\{0, x - y\}$$

(Thus, $5 \dot{-} 3 = 2$ and $3 \dot{-} 5 = 0$.) Now suppose that the underlying attribute space has dimension n. Then the probability of (O, P) is estimated to be:

$$\frac{1}{1 + \Sigma_{i=1}^{n} (P_i \dot{-} Q_i)} \tag{4}$$

where P_i and O_i are the values at the ith coordinate of the vectors P and O. To illustrate, according to (4) and Table II, the probability that deer have three layers of lipid tissue surrounding vital organs is:

$$\frac{1}{1 + ((4 \dot{-} 3) + (0 \dot{-} 1) + (5 \dot{-} 4))} = 0.33 \tag{5}$$

It is easy to see that formula (4) yields a number in $[0, 1]$ whatever vectors are associated with O, P. Probability 1 is attained if $O_i \geq P_i$ for all attributes $i \leq n$; the surplus of O_i over P_i plays no role in the calculation. Observe as well that an attribute disappears from the calculation of (4) to the extent that P_i is small; intuitively, such a value represents a nonstringent condition for possession of the predicate P. The relative salience, or importance, of an attribute is reflected in the spread of its values across objects and predicates.

An attribute like *ferocity* might be accompanied in a subject's mind by a contrasting attribute like *tameness*, and animals with high values on the former might have low values on the latter (and *vice versa*). Such pairs introduce an element of symmetry into the calculation of probability. Suppose, for example, that a given property P has the value 3 at both *ferocity* and *tameness*. Then according to (4), highly ferocious animals have little chance of possessing P (since their *tameness* value is too low) and likewise for highly tame animals (since their *ferocity* value is too low).

$P_i \dot{-} O_i$ may be conceived as the 'gap' separating object O from predicate P with respect to attribute i. Formula (4) exhibits the probability of (O, P) as a function of these gaps, hence the name 'Gap Model' for the present theory.

2.3. *Conditional Probability: One Premise Case*

Probabilities are associated with arguments like (2) in two steps. First the premises $(O^1, P) \cdots (O^m, P)$ provoke modifications in the vector representation of P, yielding a new vector P'. The statement (O, P') is then taken as the 'revised conclusion' of (2), and its probability is evaluated via formula (4). This latter probability is attributed to the argument as a whole. The transition from P to P' represents the impact of the information that our subject acquires by assuming the truth of premises $(O^1, P) \cdots (O^m, P)$. To explain the nature of this impact according to the Gap Model, we first analyze the 1-premise argument that results from suppressing the first premise of (3). It may be abbreviated as follows.

$$\frac{(\text{RABBITS, LAYERS})}{(\text{DEER, LAYERS})} \tag{6}$$

To evaluate (6) our subject must assume the truth of (RABBITS, LAYERS) and judge the probability of (DEER, LAYERS). For this purpose the Gap Model posits the following train of reasoning. Table II shows that $\text{LAYERS}_1 \dot{-} \text{RABBITS}_1 = 3$, signifying that rabbits do not have the size required of mammals with property LAYERS. However, (RABBITS, LAYERS) is a premise, hence assumed to be true. Therefore,

the property LAYERS does not require size-value 4 in order for an animal to have it. We are thus led to lower LAYERS_1, but only to the extent that rabbits resemble deer. Similarity between objects is assumed to govern the extent to which information obtained from the premise is brought to bear on the conclusion. Thus, in altering LAYERS as it applies to DEER, the Gap Model lowers LAYERS_1 by:

$$(\text{LAYERS}_1 \dot{-} \text{RABBITS}_1) \times \textit{similarity}(\text{RABBITS}, \text{DEER}) \tag{7}$$

For the similarity function in formula (7) we choose a simple measure of proximity between n-dimensional vectors, v, w, namely:

$$\textit{similarity}(v, w) = \frac{1}{1 + \textit{distance}(v, w)} \tag{8}$$

where *distance* is Euclidean distance in n-space. The range of *similarity* is seen to be $[0, 1]$. To illustrate, according to Table II:

$$\textit{similarity}(\text{RABBITS}, \text{DEER}) = \frac{1}{1 + \sqrt{(1-3)^2 + (0-1)^2 + (3-4)^2}} = 0.290$$

Thus, according to (7) the impact of the gap $\text{LAYERS}_1 \dot{-} \text{RABBITS}_1$ on the vector LAYERS is attenuated by a factor of 0.290, so only 3×0.290 is subtracted from LAYERS_1, leaving $4 - (3 \times 0.290) = 3.13$.

The gap for the second attribute, ferocity, is $\text{LAYERS}_2 \dot{-} \text{RABBITS}_2 = 0$. The second coordinate of LAYERS is therefore reduced by $0 \times$ *similarity*(RABBITS, DEER) and so retains its original value.

Finally, LAYERS_3 is reduced by

$$(\text{LAYERS}_3 \dot{-} \text{RABBITS}_3) \times \textit{similarity}(\text{RABBITS}, \text{DEER}) = (5 \dot{-} 3) \times 0.290 = 0.580$$

and becomes $\text{LAYERS}_3 - 0.580 = 5 - 0.580 = 4.42$.

The premise (RABBITS, LAYERS) of argument (6) has thus modified the vector LAYERS from its original state shown in Table II to the new values $\text{LAYERS}' = (3.13, 0, 4.42)$. It remains only to calculate the probability of (DEER, LAYERS') according to formula (4). This yields:

$$\frac{1}{1+(3.13\div 3)+(0\div 1)+(4.42\div 4)} = \frac{1}{1+0.13+0+0.42}$$
$$= 0.65\,.$$

Observe that the latter probability exceeds the unconditional probability of (DEER, LAYERS) computed in (5). The difference is due to the impact of the premise (RABBITS, LAYERS), which changes our subject's interpretation of LAYERS, bringing it into greater conformity with the vector underlying DEER.

Intuitively, a statement (O, P) that gives rise to large gaps $P_i \div O_i$ is implausible, since O fails to meet conditions embodied in P. By the same token, such a statement constitutes a surprising premise, and thus tends to raise the probability of associated conclusions.[1] The dual role of gaps is represented by (4) for statements and by gap-reduction as discussed above for premises. The impact of premise gap is modulated in our theory by multiplication with the similarity obtained between premise and conclusion categories. Greater similarity is thus assumed to increase the relevance to the conclusion of the information contained in the premise.[2]

2.4. Conditional Probability: Multiple Premises

The Gap Model's analysis of multiple-premise arguments is motivated by the following principle.

PRINCIPLE OF PREMISE DIVERSITY. Adding a new premise (O^{m+1}, P) to an argument

$$\begin{array}{l}(O^1, P)\\ \vdots\\ \underline{(O^m, P)}\\ (O, P)\end{array} \qquad (9)$$

raises the probability of (O, P) only to the extent that O^{m+1} differs from $O^1 \cdots O^m$.

Documentation of (9) in human reasoning may be found in [23].[3] The Gap Model takes account of premise diversity by a maximum-principle for calculating the impact of multiple premises on the predicate vector.

We may illustrate with argument (3), abbreviated to:

$$\begin{array}{l}(\text{LIONS, LAYERS}) \\ \underline{(\text{RABBITS, LAYERS})} \\ (\text{DEER, LAYERS})\end{array} \qquad (10)$$

The *potential impact* of the premise (RABBITS, LAYERS) on the size-value of LAYERS is defined by (7), yielding $3 \times 0.290 = 0.870$. Likewise, the potential impact of (LIONS, LAYERS) on the size-value of LAYERS is

$$(\text{LAYERS}_1 \dot{-} \text{LIONS}_1) \times \textit{similarity}(\text{LIONS, DEER}) .$$

This number is:

$$(4 \dot{-} 3) \times \frac{1}{1 + \sqrt{0^2 + 5^2 + 7^2}} = 0.104 .$$

Since the potential impact on size of (RABBITS, LAYERS) exceeds that of (LIONS, LAYERS), the size-value of LAYERS is decreased by the former rather than by the latter. Hence LAYERS_1 declines by 0.870 to 3.13.

On the psychological level, the (assumed) fact that rabbits have three layers of lipid tissue surrounding vital organs provides more information about the minimal size required for deer to possess this property than does the corresponding fact about lions. Indeed, according to Table II (LIONS, LAYERS) provides little information since lions are already assumed to have nearly the required size; additionally, lions have low resemblance to deer. On the other hand, (RABBITS, LAYERS) is quite informative since rabbits have much less of the size previously thought to be necessary; additionally they resemble deer more than lions do (once again, according to the table).

The third attribute provides a contrasting case. The potential impact of (LIONS, LAYERS) on *frigid-habitat* equals

$$(\text{LAYERS}_3 \dot{-} \text{LIONS}_3) \times \textit{similarity}(\text{LIONS}, \text{DEER}) = 8 \times 0.104 = 0.832 .$$

This exceeds the potential impact of (RABBITS, LAYERS) on *frigid-habitat*, which is:

$$(\text{LAYERS}_3 \dot{-} \text{RABBITS}_3) \times \textit{similarity}(\text{RABBITS}, \text{DEER}) = 2 \times 0.290 = 0.580 .$$

Hence, it is the gap provoked by (LIONS, LAYERS) and attenuated by *similarity*(LIONS, DEER) that decreases LAYERS_3.

The foregoing process yields a modified predicate-vector LAYERS′. The probability associated with argument (10) is then computed as before from formula (4). On the basis of Table II this number is 0.77, which is higher than for the 1-premise argument (6).

Our use of maximization ensures that an argument $(s \mid s_1, s_2)$ whose premises bear on highly similar objects will be assigned roughly the same probability as $(s \mid s_1)$. In contrast, if s_1, s_2 involve dissimilar objects, then the potential impact induced by a given attribute has an additional chance to exceed its potential impact in $(s \mid s_1)$. Diversity of premises thus tends to increase the probability of $(s \mid s_1, s_2)$ compared to $(s \mid s_1)$. In this sense maximization implements principle (9).

2.5. *Summary of the Model*

Operative formulas. Let A, B, and P be real vectors of length n (conceived as two object vectors and a predicate vector, respectively).

$$\textit{prob}(A, P) = \frac{1}{1 + \Sigma_{i=1}^{n} (P_i \dot{-} A_i)} \tag{11}$$

$$\textit{similarity}(A, B) = \frac{1}{1 + \textit{distance}(A, B)} \tag{12}$$

(where *distance* is Euclidean distance).

For the next formula we conceive of (A, P) as a premise and (B, P) as the conclusion of a given argument. Let $i \leqslant n$ be given.

$$potential\ impact(A, B, P, i) = (P_i \dot{-} A_i) \times similarity(A, B) \tag{13}$$

Evaluation of arguments. Let argument

$$\begin{array}{l} (O^1, P) \\ \vdots \\ \underline{(O^m, P)} \\ (O, P) \end{array}$$

be given, and suppose that its objects and predicate are represented by real vectors of length n. The (conditional) probability associated with this argument is calculated as follows.

If $m = 0$ then the probability is $prob(O, P)$.

If $m \geqslant 1$ then the probability is $prob(O, P')$, where P' is the length n vector whose ith coordinate is calculated as follows:

$$\begin{aligned} &P'_i = P_i - potential\ impact(A, O, P, i), \text{ where} \\ &\quad A \in \{O^1 \cdots O^m\} \quad \text{and} \\ &potential\ impact(A, O, P, i) \geqslant potential\ impact(B, O, P, i) \\ &\quad \text{for all } B \in \{O^1 \cdots O^m\}\,. \end{aligned} \tag{14}$$

2.6. *Alternative Realizations of the Gap Model*

The Gap Model rests on five psychological hypotheses, which may be formulated as follows.

(a) The mental representation of objects and predicates can in large part be summarized by real vectors in an appropriate attribute space.

(b) A statement (O, P) is perceived to be probable to the extent that attributes evoked by the predicate are present in the object.
(c) An argument's premises increase the probability of its conclusion by lowering the attribute values presumed necessary for possession of the property in question.
(d) The impact of a premise depends on (i) the disparity between its attribute-levels and those of the predicate, and (ii) the similarity of the premise-object to the conclusion-object.
(e) The impact of multiple premises is governed by the maximum principle of Section 2.4 (which entails, in practice, the diversity principle (9)).

The formulas of our model realize hypotheses (a)–(e) in an extremely simple way, and alternatives naturally come to mind. For example, the arbitrary constant '1' in both (11) and (12) could be replaced by larger constants in order to decelerate the descent of these functions towards zero. Or, (11) might be replaced by

$$prob(A, P) = \frac{1}{1 + \frac{1}{n} \sum_{i=1}^{n} (P_i \dot{-} A_i)} \tag{15}$$

where n is the dimension of the underlying attribute space. The advantage of (15) is that adding dimensions to a space would not have a tendency to lower the probabilities of statements since probability according to (15) depends on average gaps rather than on their sum. A similar modification could be made to (12).

As another example, the similarity function chosen for (12) is based on the Euclidean metric, but it is well known that human similarity judgment violates the metric axioms (see [36, 37]). The Contrast Model of similarity ([36, 21]) is known to have greater psychological fidelity in this regard, and has figured in other studies of probability judgment (e.g. [24, 35]). It might thus be usefully substituted for the function defined in (12).

As a third example, our maximum principle implies that the probability of a conclusion monotonically increases with additions to the premise-set. However, counterexamples to monotonicity in human judgment have been demonstrated in [23, 30]. Principles of 'coverage'

discussed in [23] could be adjoined to the maximum principle in order to account for such counterexamples.

Finally, a more sophisticated elaboration of the model would supplement its vectorial representation of knowledge with the frame-based architecture discussed in [1, 18, 19]. Evidence for the psychological reality of frames is considered in [28, 34].

We have investigated a variety of proposals like these, applying elaborated versions of the Gap Model to the data described below. It will be more revealing here, however, to retain the simple version of the model, as summarized in Section 2.5. The positive results obtained for extrapolation will then be more easily interpreted as favorable to the general approach we advocate.

In more general terms, it is not the purpose of the present paper to insist on the particular choices embodied by the Gap Model. It is enough at present to show that solution to the extrapolation problem can at least be envisioned, and to document some predictive success for a simple model.

2.7. Normative Status of the Gap Model

Numerical analysis of the Gap Model suggests that it assigns probabilities to arguments in a manner consistent with the standard probability axioms. This feature of the model, however, is an artifact of the restricted form of arguments to which it currently applies. When the model is extended to a broader class of arguments, it becomes normatively deviant. Minor modifications of the current version also bring it into conflict with the probability axioms. This results, for example, if probability function (11) is replaced by (15) (we omit the details).

Fidelity to the probability calculus is a mixed blessing for models of the kind at issue in this paper. It is an advantage for applications to objective probabilities (generated, for example, by a database). In contrast, it can be an impediment to modeling human judgment, which in many contexts does not strictly adhere to the probability axioms ([12, 38, 30]). Indeed, it will be seen that subjects in our extrapolation experiments sometimes violated a simple law of probability.

Finally, it is worth pointing out that no normative theory currently

exists relating similarities among categories to the probabilities of statements. People nonetheless exploit similarity for just this purpose on a daily basis (see [31] for discussion). It may be hoped that reflection on the extrapolation problem will generate insight into the normative role of similarity in probability assessment.

3. EXTRAPOLATION EXPERIMENT

Our experiment was performed on 52 undergraduates from the University of Michigan, recruited by advertisement and paid for their participation.

3.1. Method

The experimental protocol may be divided into three parts. First, subjects were presented with a set of objects and predicates. Second, they assigned probabilities to every argument of form (2) induced by the set. Third, as a reliability check, they made the same judgments a second time – on the same arguments but in a different random order. We consider these parts in turn.

Presentation of objects and predicates. Subjects were randomly assigned one of the two sets of stimuli in Table III. A stimulus set

TABLE III

Sets of objects and predicates available as options.

	Set 1
Objects:	Bears, Beavers, Squirrels, Monkeys, Gorillas
Predicate 1:	have 3 distinct layers of fat tissue surrounding vital organs
Predicate 2:	have over 80% of their brain surface devoted to neocortex
	Set 2
Objects:	Lions, Housecats, Camels, Elephants, Hippos
Predicate 1:	have a visual system that fully adapts to darkness in less than 5 minutes
Predicate 2:	have skins more resistant to penetration than most synthetic fibers

consisted of five mammal species to serve as objects plus two predicates. It was verified for each subject that the mammals of the chosen option were familiar and easily distinguishable, and that the chosen predicates were interpretable and meaningful. The sets were constructed so as to manifest a range of similarity among the five mammals (at least, in the judgment of the experimenters).

The five objects and either one of the two predicates assigned to the subject generate 80 arguments, as described in Table I. Since only one predicate may appear in a given argument, these two sets of 80 arguments exhaust the possibilities.

Assignment of probabilities. Each subject assigned probabilities to his 160 arguments, delivered in random order via computer. For multi-premise arguments, the order of premises was determined randomly. To illustrate, a typical 2-premise argument was presented as follows.

```
What is the probability that
  Bears have over 80% of brain surface
  devoted to neocortex
given that this is true of:
  squirrels and beavers.
Probability: ___
```

The 'given that' clause did not appear for 0-premise arguments. In prior instructions it was emphasized that probabilities must be assigned while assuming the truth of given premises (if any). On the other hand, each question was to be treated separately, with no assumptions carried forward.

The first two parts of the procedure were performed in immediate succession, and required roughly one hour to complete.

Reliability check. One to three days later subjects returned to evaluate their 160 arguments for a second time. The arguments were delivered in a new random order; premise-order within multi-premise arguments was also freshly randomized. The subject's previous responses were not made available to him.

3.2. *Preliminary Analyses*

For each subject the Pearson correlation was computed between his responses to corresponding arguments in parts 2 and 3 of the procedure. This correlation was less than 0.7 for 22 of the subjects, who were dropped from all further analyses. The median reliability for the remaining 30 subjects is 0.80. In all ensuing analysis we use the average of a subject's two responses to the same argument as the 'official' probability he assigns to that argument.

The following analysis indicates the degree to which the judgments of our 30 subjects deviate from the probability calculus. It is well known that for any two statements p, q the axioms of probability require:

$$\Pr(p \wedge q) \geqslant \Pr(p) + \Pr(q) - 1 \tag{16}$$

Since $\Pr(p \mid q) = \Pr(p \wedge q)/\Pr(q)$, (16) implies:

$$\Pr(p \mid q) \times \Pr(q) \geqslant \Pr(p) + \Pr(q) - 1 \tag{17}$$

Each subject evaluated forty 1-premise arguments of form $(p \mid q)$ along with the corresponding statements p, q. Hence, each subject had forty occasions to violate inequality (17). In fact, 22 of the 30 subjects violated (17) at least once. The average number of violations over all 30 subjects is 6.7.

3.3. *Extrapolation Based on the Gap Model*

Extrapolation analyses using the Gap Model were performed on a within-subject basis via the following five steps.

Step 1. The 160 arguments evaluated by a given subject were segregated into two sets of 80 according to the predicate appearing therein. Each set of 80 arguments was treated separately, thereby dividing each of the thirty subjects into two halves. In the sequel we shall refer to these 60 data-sets (two per subject) as 'half-subjects'.

Step 2. The 80 arguments of a given, half-subject were partitioned into two sets. One set was used to fix the parameters of the Gap Model

TABLE IV

Extrapolation using correlation as a measure of fit. Columns 2 and 3 describe the arguments used to fix the parameters of the model, and those predicted subsequently. Columns 4–6 present the median correlations obtained in the testing phase for the original Gap Model, the model with similarity set uniformly to 1.0, and the model with the maximum-principle replaced by addition. The medians are computed over sixty, half-subjects.

	Arguments		Median Correlation		
	used for fixing	*used for testing*	*Gap*	*NoSim*	*NoMax*
1)	60 non-3-premise	20 3-premise	0.88	0.74	0.82
2)	60 random	remaining 20	0.85	0.76	0.86
3)	50 non-2-premise	30 2-premise	0.88	0.75	0.85
4)	50 random	remaining 30	0.84	0.77	0.85
5)	30 2-premise	50 non-2-premise	0.79	0.70	0.79
6)	30 random	remaining 50	0.77	0.70	0.80
7)	20 3-premise	60 non-3-premise	0.73	0.49	0.65
8)	20 random	remaining 60	0.72	0.60	0.76

(as described below); the other set tested the predictions of the model once its parameters were fixed. Eight kinds of partitions were employed, listed in Table IV. For example, row (1) of Table IV refers to the partition in which the 60 arguments were chosen randomly for parameter-fixing, and the remaining 20 were used for testing. Such random partitions were generated afresh for each of the 60 half-subjects.

Step 3. A dimensionality n for the underlying attribute space was chosen. In the example of Section 2, $n = 3$. For the present analyses we used both $n = 2$ and $n = 3$. Small values of n are suggested by multidimensional scaling solutions to judgments of similarity among members of natural categories like mammals or birds. Typically, two or three dimensions suffice to approximately represent such judgements in euclidean space (see [27, 2]). For brevity, we discuss only the choice $n = 2$; the results for $n = 3$ are entirely comparable. Thus, in what follows we assume that the five objects and one predicate appearing in the 80 arguments of a half-subject are each associated with real values on two attributes. Twelve $(= 2(5 + 1))$ parameters must therefore be fixed in order for the Gap Model to make predictions about new arguments. No attempt was made to identify the two attributes (*size*, *ferocity*, etc.) presumed to underlie subjects'

representations of objects and predicates. These attributes are simply formal place-holders in what follows.

Step 4. For each half-subject, an iterative procedure was employed to find values of the 12 free parameters that maximize the Gap Model's fit to the initial, fixing set of arguments (the testing arguments play no role in this step). To illustrate, consider the partition described by the first row of Table IV. Choice of the 12 parameters causes the Gap Model to assign probabilities to each of the sixty arguments with $0, 1, 2$, or 4 premises. These sixty probabilities may be compared to those selected by a given subject. As a measure of goodness-of-fit, we calculated the Pearson correlation between the corresponding probabilities assigned by model and subject to each of these 60 arguments. (A different measure of fit is discussed below.) The set of 12 parameters that maximize this correlation was retained. Independent maximization was carried out for each of the sixty, half-subjects with respect to each of the eight partitions in Table IV (480 maximizations in all). The maximization algorithm employed was based on the 'direction set' method described in [26, Chapter 10]. Twenty starting points were tried, chosen uniformly-randomly within $[-2, 2]^{12}$. The best set of parameters over all twenty runs was retained.

Step 5. Once the best set of 12 parameters – associated with a given, half-subject and a given partition of arguments – was obtained, the Gap Model with those parameters was applied to the 'testing' arguments of the partition in question. The probabilities generated by the model in this way were then compared to the corresponding probabilities assigned by the subject. The Pearson correlation between these sets of probabilities was used as a measure of fit.

For each of the eight partitions, the column headed '*Gap*' in Table IV shows the median correlation obtained in step 5 over all 60 half-subjects.[4] Note that even when the information fed into our extrapolation algorithm is limited to 20, randomly chosen arguments, it predicts the remaining 60 arguments up to a correlation of 0.72. Figure 1 shows the scatter plot for the half-subject whose correlation is at the median value 0.88 with respect to the partition described in the first row of Table IV.[5] Likewise, Figure 2 shows the scatter plot for the half-subject at the median value with respect to partition 2.

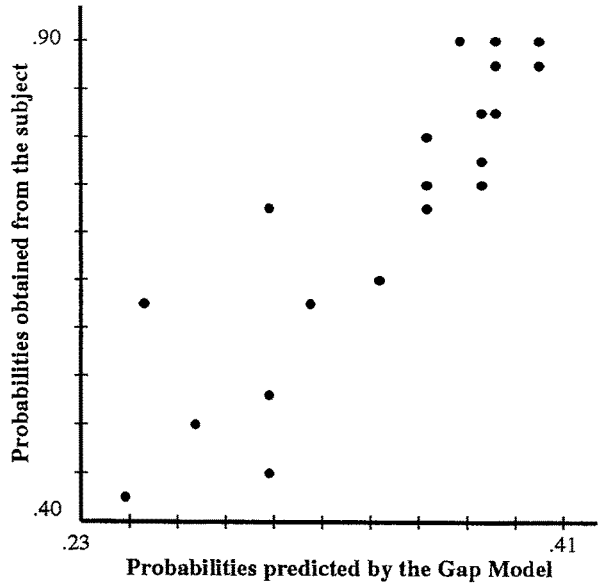

Fig. 1. Plot for partition 1 of Table IV.

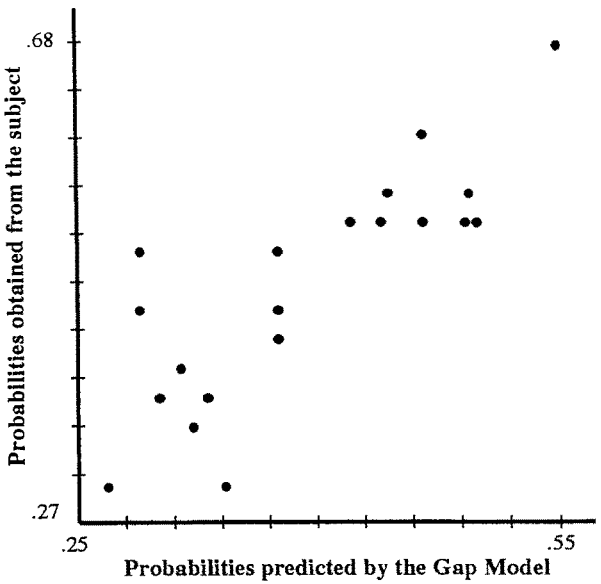

Fig. 2. Plot for partition 2 of Table IV.

3.4. Extrapolation Based on Variants of the Gap Model

As a further test of the psychological reality of the Gap Model, we considered two variant models differing from the original in selected ways. The first variant removed considerations of similarity from the Gap Model by uniformly imposing the value 1.0 as the outcome of all similarity calculations; in other respects the first variant is the same as the original. Thus, the new model – called *NoSim* – results from replacing formula (12) by:

$$similarity(A, B) = 1.0 \qquad \text{for all objects } A, B\,.$$

The second, variant model results from replacing the maximum principle (14) by a principle of addition, formulated as follows. Let an m-premise argument

$$\begin{array}{l} (O^1, P) \\ \vdots \\ \underline{(O^m, P)} \\ (O, P) \end{array}$$

with $m \geqslant 1$ be given. The probability assigned to this argument is $prob(O, P')$ where:

$$P'_i = P_i - \sum_{j=1}^{m} potential\ impact(O^j, O, P, i)$$

(The function *potential impact* is defined in formula (13).) Thus, the present variant – called *NoMax* – lowers the values of P by the total impact of the premises, rather than by their maximum.

The same extrapolation analyses performed on the Gap Model were repeated on these two variants. The results are summarized in columns 5 and 6 of Table IV. The Gap Model performs better than *NoSim*, thus giving indirect support to the role of similarity in probability judgment. The comparison with *NoMax*, however, offers little evidence for the

TABLE V

Direct comparison of the Gap Model and its two variants, using correlation as a measure of fit. Columns 2 and 3 summarize the partitions used. Column 4 shows the number of half-subjects (out of 60) whose correlation in step 5 favors the Gap Model over *NoSim*. Column 5 provides the same information with respect to *NoMax*.

	Arguments		Comparison of models	
	used for fixing	*used for testing*	*Gap* vs. *NoSim*	*Gap* vs. *NoMax*
1)	60 non-3-premise	20 3-premise	50	41
2)	60 random	remaining 20	37	38
3)	50 non-2-premise	30 2-premise	49	37
4)	50 random	remaining 30	41	32
5)	30 2-premise	50 non-2-premise	46	33
6)	30 random	remaining 50	46	30
7)	20 3-premise	60 non-3-premise	51	39
8)	20 random	remaining 60	41	28

maximum principle. Greater support emerges from the finer analysis summarized in Table V. Column 4 of the table shows the number of half-subjects (out of 60) for whom the Gap Model provides a higher correlation in step 5 than does *NoSim*. Column 5 provides the same comparison with respect to *NoMax*.

3.5. Absolute Deviation as a Measure of Fit

Instead of maximizing correlation coefficients, as above, another natural measure of fit between model and data is the average, absolute deviation between predicted and observed probabilities. To illustrate, consider again the first partition of Table IV. Fixing the Gap Model's 12 free parameters causes it to assign probabilities p_i to each of the sixty arguments with 0, 1, 2, or 4 premises. A given subject assigns probabilities q_i to the same arguments. As a measure of fit, we now use the average over $|p_i - q_i|$, instead of correlation. The same measure of fit is used in step 5, to test the model. Of course, we now seek to *minimize* the average absolute deviation, compared to *maximizing* the correlation. Otherwise the details of the optimization procedure are the same.

All of the preceding analyses were repeated using absolute deviation in place of correlation. The results are summarized in Tables VI and VII. To illustrate, the number '0.075' in row 1, column 4 of Table VI

TABLE VI

Extrapolation using absolute deviation as a measure of fit. Columns 2 and 3 describe the arguments used to fix the parameters of the model, and those predicted subsequently. Columns 4–6 present the median, average, absolute deviation obtained in the testing phase for the original Gap Model, the model with similarity set uniformly to 1.0, and the model with the maximum-principle replaced by addition. The medians are computed over sixty, half-subjects.

	Arguments		Median Deviation		
	used for fixing	*used for testing*	*Gap*	*NoSim*	*NoMax*
1)	60 non-3-premise	20 3-premise	0.075	0.169	0.087
2)	60 random	remaining 20	0.080	0.174	0.094
3)	50 non-2-premise	30 2-premise	0.074	0.174	0.085
4)	50 random	remaining 30	0.084	0.173	0.098
5)	30 2-premise	50 non-2-premise	0.096	0.190	0.109
6)	30 random	remaining 50	0.102	0.188	0.118
7)	20 3-premise	60 non-3-premise	0.132	0.267	0.144
8)	20 random	remaining 60	0.118	0.203	0.130

TABLE VII

Direct comparison of the Gap Model and its two variants, using absolute deviation as a measure of fit. Columns 2 and 3 summarize the partitions used. Column 4 shows the number of half-subjects (out of 60) whose average, absolute deviation in step 5 favors the Gap Model over *NoSim*. Column 5 provides the same information with respect to *NoMax*.

	Arguments		Comparison of models	
	used for fixing	*used for testing*	*Gap* vs. *NoSim*	*Gap* vs. *NoMax*
1)	60 non-3-premise	20 3-premise	60	34
2)	60 random	remaining 20	54	31
3)	50 non-2-premise	30 2-premise	57	35
4)	50 random	remaining 30	55	32
5)	30 2-premise	50 non-2-premise	60	39
6)	30 random	remaining 50	59	38
7)	20 3-premise	60 non-3-premise	59	39
8)	20 random	remaining 60	52	37

indicates the absolute size of the error committed by the Gap Model when it is used to extrapolate 3-premise arguments from the rest. In terms of absolute deviation, the Gap Model is seen to outperform both the *NoSim* and *NoMax* variants.

We note that maximizing the correlation in Step 4 of our extrapolation analysis tends not to minimize the median, absolute deviation in

Step 5; nor does minimizing the latter maximize the former. It thus appears necessary to choose in advance the desired kind of extrapolation.

4. CONCLUDING REMARKS

Despite its simplicity, the Gap Model enjoys nonnegligible success in extrapolation. We interpret this result as encouraging the view that successors to the Gap Model could eventually provide reasonably accurate models of probability judgment in natural domains of reasoning.

The practical interest of such models is highlighted by recent progress in the theory of influence diagrams [8] and belief nets [25, 14] (see [20, 29] for an introduction). This work provides a set of tools for constructing efficient systems of decision-making and analysis that are grounded in the theory of utility and probability. Use of the tools, however, often requires large numbers of conditional probabilities to be elicited from an external, human agent (for example, many thousands in the systems built by Heckerman [4]). A successful method of extrapolation might allow fewer judgments to be elicited; the remaining judgments would be estimated. Likewise, a small set of missing probabilities – unforeseen at the outset of a project – could be extrapolated at a later stage from stored probabilities.

Extrapolation might also be used to enlarge the set of conditional probabilities that can be estimated from a database. To explain, suppose that data are available about the occurrence and co-occurrence of binary categories $A_1 \cdots A_n$. The data might be numerous enough to empirically estimate conditional probabilities of form $\Pr(A_i | A_j)$ but not of forms $\Pr(A_i | A_j, A_k)$, $\Pr(A_i | A_j, A_k, A_l)$, etc. This situation will occur whenever the number of categories A_i is too large for the number of records in the database, since complex conditioning events will occur too infrequently to allow meaningful estimates of the probability of their subevents.

Extrapolation might nonetheless provide a subjectively plausible guess about the missing probabilities. This would be achieved by the use of a 'dummy' predicate P asserting that a record drawn randomly from the database falls into whatever category is associated with it.

Thus (A, P) is the statement that a given record falls into category A, and the probability attached to this statement can be estimated directly from the database. Similarly, the argument

$$\frac{(A_1, P)}{(A_2, P)}$$

represents the proportion of records in category A_2 among those in category A_1. These numbers in hand, an extrapolation algorithm provides conjectures about the conditional probabilities embodied in arguments of arbitrary complexity. Application of the Gap Model, for example, would proceed by seeking featural representations for the categories at issue, as well as for the dummy P. The features sought for P would be those representing a typical or modal record in the database, giving rise thereby to appropriate gaps with respect to the categories $A_1 \ldots A_n$. Whatever the extrapolation algorithm employed, if it is based on an adequate psychological theory, its conjectures will enjoy the plausibility of human judgment. The objective accuracy of these judgments can then be compared to those delivered by more familiar principles of 'ampliative inference' such as maximizing entropy (see [10, 11, 13, 22]).[6]

More fundamentally, the kind of model envisioned in this paper would be able to convert information about object- and predicate-attributes into conditional probabilities of arguments. Suppose, for example, that a database contained 100 objects and 100 predicates, each coded along five attributes; it would thus contain $5 \times (100 + 100)$ or 1000 values. In contrast, 100 objects and 100 predicates generate an astronomical number of arguments, any of whose probabilities might be needed in an associated system of automated reasoning. Construction of the reasoning system would be facilitated by an algorithm that could examine the available database and supply reasonable approximations to the probabilities a human agent would attribute to the arguments in question. Such an algorithm would be particularly useful in any attempt to automate the synthesis of reasoning systems whose performance need not exceed the standard of common sense.

ACKNOWLEDGEMENTS

Research support was provided by Swiss National Science Foundation Contract No. 21-32399.91 to Osherson, Air Force Contract No.

AFSOR-91-0265 to Smith and by U.S. Public Health Service Grant No. 1-R29-MH46885 to Shafir. We thank Antoine Gualtierotti and an anonymous reviewer for helpful comments. Correspondence to D. Osherson, IDIAP, C.P. 609, CH-1920 Martigny, Switzerland; e-mail: osherson@maya.idiap.ch.

NOTES

[1] For discussion of surprise in evaluating the probability of arguments, see [9, Chapter 4].

[2] The role of similarity in induction is reviewed in [31]. Evidence that similarity is evaluated along multiple dimensions within reasoning tasks is presented in [5].

[3] The plausibility of (9) from a normative, epistemological point of view is discussed in [6, 7].

[4] As expected, the correlations obtained in the retrodictive step 4 are systematically higher than those obtained in prediction.

[5] Because there are an even number of half-subjects, the median value is actually straddled by two subjects. We show the plot for the lower subject.

[6] Note that entropy principles cannot be applied to our psychological data because the latter are not consistent with the probability calculus (see Section 3.2).

REFERENCES

1. Bobrow, D. and Winograd, T.: 1977, 'An overview of KRL, a knowledge representation language, *Cognitive Science* **1**(1), 3–46.
2. Caramazza, A., Hersch, H. and Torgerson, W.: 1976, 'Subjective structures and operations in semantic memory', *Journal of Verbal Learning and Verbal Behavior* **15**, 103–118.
3. Collins, A. and Loftus, E.: 1975, 'A spreading activation theory of semantic processing', *Psychological Review* **82**, 407–428.
4. Heckerman, D.: 1990, *Probabilistic Similarity Networks*, Dissertation submitted to the program in Medical Information Sciences, Stanford University.
5. Heit, E. and Rubinstein, J.: 1992, 'Similarity and property effects in inductive reasoning', MS, University of Michigan, Department of Psychology.
6. Hempel, C.G.: 1966, *Philosophy of Natural Science*, Englewood Cliffs, NJ: Prentice-Hall.
7. Horwich, P.: 1982, *Probability and Evidence*, New York: Cambridge University Press.
8. Howard, R. and Matheson, J.: 1984, 'Influence diagrams', *In* Howard R., and Matheson, J. (Eds.): 1984, *Applications of Decision Analysis*, Vol II, Strategic Decisions Group, Menlo Park, California.
9. Howson, C. and Urbach, P.: 1989, *Scientific Reasoning: The Bayesian Approach*, La Salle, IL: Open Court.
10. Hunter, D.: 1986, 'Uncertain reasoning using maximum entropy inference; *In*

Kanal, L. and Lemmer, J. (Eds.): 1986, *Uncertainty in Artificial Intelligence*, Amsterdam: Elsevier.
11. Jaynes, E.: 1979, 'Where do we stand on maximum entropy? *In* Levine, R. and Tribus, M. (Eds.): 1979, *The Maximum Entropy Formalism*, Cambridge: M.I.T. Press.
12. Kahneman, D., Slovic, P. and Tversky, A.: 1980, *Judgment Under Uncertainty: Heuristics and Biases*, New York: Cambridge University Press.
13. Klir, G.: 1988, 'Methodological principles of uncertainty in inductive modeling: A new perspective, *In* Erickson, G. and Smith, C. (Eds.): 1988, *Maximum Entropy and Bayesian Methods in Science and Engineering* Vol. 1, Dordrecht: Kluwer Academic Publishers.
14. Lauritzen, S. and Spiegelhalter, D.: 1988, 'Local computations with probabilities on graphical structures and their applications to expert systems', *Journal of the Royal Statistical Society B*, **50**(2).
15. Malt, B. and Smith, E.E.: 1984, 'Correlated properties in natural categories', *Journal of Verbal Learning and Verbal Behavior* **23**, 250–269.
16. Medin, D., Altom, M., Edelson, S., and Freko, D.: 1982, 'Correlated symptoms and simulated medical diagnosis, *Journal of Experimental Psychology: Learning Memory, and Cognition* **8**, 37–50.
17. Medin, D.L. and Shoben, E.J.: 1988, 'Context and structure in conceptual combination; *Cognitive Psychology* **20**, 158–190.
18. Minsky, M.: 1981, 'A framework for representing knowledge', *In* Haugeland, J. (Ed.): 1981, *Mind Design*, Cambridge, MA: M.I.T. Press, pp. 95–128.
19. Minsky, M.: 1986, *The Society of Mind*, New York: Simon & Schuster.
20. Neapolitan, R.: 1990, *Probabilistic Reasoning in Expert Systems*, New York: Wiley.
21. Osherson, D.: 1987, 'New axioms for the contrast model of similarity', *Journal of Mathematical Psychology* **31**(1), 93–103.
22. Osherson, D., Shafir, E. and Smith, E.: (in press) 'Ampliative inference: On choosing a probability distribution, *Cognition*.
23. Osherson, D., Smith, E., Wilkie, O., López, A. and Shafir, E.: 1990, 'Category based induction, *Psychological Review* **97**(2), 185–200.
24. Osherson, D., Stern, J., Wilkie, O., Stob, M. and Smith, E.E.: 1991, 'Default probability, *Cognitive Science* **15**, 251–270.
25. Pearl, J.: 1988, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, San Mateo, CA: Morgan-Kaufmann.
26. Press, W.H., Flannery, B.P., Teukolsky, S.A., and Vetterling, W.: 1988, *Numerical Recipes in C*, New York: Cambridge University Press.
27. Rips, L., Shoben, E. and Smith, E.E.: 1973, 'Semantic distance and the verification of semantic relations', *Journal of Verbal Learning and Verbal Behavior* **12**, 1–20.
28. Rumelhart, D. and Ortony, A.: 1977, 'The representation of knowledge in memory', *In* Anderson, R., Spiro, R. and Montague, W. (Eds.): 1977, *Schooling and the Acquisition of Knowledge*, Hillsdale, NJ: Erlbaum.
29. Shafer, G. and Pearl, J. (Eds.): 1990, *Readings in Uncertain Reasoning*, San Mateo, CA: Morgan-Kauffmann.
30. Shafir, E., Smith, E.E. and Osherson, D.: 1990, 'Typicality and reasoning fallacies', *Memory and Cognition* **18**(3), 229–239.
31. Smith, E.E., Shafir, E. and Osherson, D.: (in press), 'Similarity, plausibility, and judgments of probability', *Cognition*.

32. Smith, E.E. and Medin, D.: 1981, *Categories and Concepts*, Cambridge, MA: Harvard University Press.
33. Smith, E.E., Osherson, D.N. Rips, L.J. and Keane, M.: 1988, 'Combining prototypes: A selective modification model', *Cognitive Science* **12**, 485–527.
34. Smith, E.E.: 1989, 'Concepts and induction', *In* Posner, M. (Eds.): 1989, *Foundations of Cognitive Science*, Cambridge, MA: M.I.T. Press, pp. 502–526.
35. Stern, Joshua: 1991, *Similarity-Based Likelihood Judgment*, M.I.T. Dissertation, Department of Brain and Cognitive Sciences.
36. Tversky, A.: 1977, 'Features of similarity', *Psychological Review* **84**(4), 327–362.
37. Tversky, A. and Gati, I.: 1982, 'Similarity, separability, and the triangle inequality', *Psychological Review* **89**(2), 123–154.
38. Tversky, A. and Kahneman, D.: 1983, 'Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment', *Psychological Review* **90**, 293–315.

Daniel Osherson
IDIAP,
1920 Martigny,
Case Postale 609,
Valais, Switzerland

Edward E. Smith
University of Michigan

Tracy S. Myers
M.I.T.

Eldar Shafir
Princeton University

Michael Stob
Calvin College