Cognition, 49 (1993) 67–96
0010-0277/93/$06.00 © 1993 – Elsevier Science Publishers B.V. All rights reserved.

Similarity, plausibility, and judgments of probability

Edward E. Smith*,[a], Eldar Shafir[b], Daniel Osherson[c]

[a] *Department of Psychology, University of Michigan, 330 Packard Road, Ann Arbor, MI 48104, USA*
[b] *Department of Psychology, Princeton University, Princeton, NJ 08544, USA*
[c] *IDIAP, CP 609, 1920 Martigny, Valais, Switzerland*

Abstract

Judging the strength of an argument may underlie many reasoning and decision-making tasks. In this article, we focus on "category-based" arguments, in which the premises and conclusion are of the form All members of C have property P, *where C is a natural category. An example is "Dobermanns have sesamoid bones. Therefore, German shepherds have sesamoid bones." The strength of such an argument is reflected in the judged probability that the conclusion is true given that the premises are true. The processes that mediate such probability judgments depend on whether the predicate is "blank" – an unfamiliar property that does not enter the reasoning process (e.g., "have sesamoid bones") – or "non-blank" – a relatively familiar property that is easier to reason from (e.g., "can bite through wire"). With blank predicates, probability judgments are based on similarity relations between the premise and conclusion categories. With non-blank predicates, probability judgements are based on both similarity relations and the plausibility of premises and conclusion.*

Introduction

Reasoning and decision making in the face of uncertainty often require one to estimate the probabilities of uncertain events. In a series of influential studies,

The research reported in this article was supported by Air Force Contract No. AFOSR-92-0265 to Smith, US Public Health Service Grant No. 1-R29-MH46885 to Shafir, and Swiss National Science Foundation Contract No. 21-32399.91 to Osherson. We thank Douglas Medin for helpful comments on an earlier version of the manuscript.

*Corresponding author.

Kahneman and Tversky (e.g., 1973, Tversky & Kahneman, 1983) demonstrated that lay people base their intuitive estimates of probability on decision heuristics, which, though often useful, sometimes yield normatively incorrect judgments. One such heuristic estimates the probability that individual *i* has property *P* in terms of how representative *i* is of *P*. Many empirical studies of this heuristic have involved a paradigm in which subjects are presented with a description of a hypothetical person, and asked to estimate the probability that the person is an instance of a target category; for example, "Linda is 31, liberal, and outspoken. What is the probability that she is a social worker?" In cases like this, the representativeness of the individual reduces to the typicality of the instance in the target-category – roughly, how good an example the instance is of the category – and the critical finding is that probability judgments are an increasing function of typicality (Shafir, Smith, & Osherson, 1990).

The Kahneman–Tversky paradigm bears on contexts in which one needs to estimate the probability that an object belongs to a particular category. There is, however, another natural paradigm in which instances and categories are used to support probability judgments. In this paradigm, subjects are informed that some members of a category have a particular property, and then have to estimate the probability that other members have the property as well; for example, "A majority of surgeons oppose socialized medicine. What is the probability that a majority of internists do so as well?" These inferences are said to be "category based". Presumably, subjects are treating surgeons and internists as subsets of the category of physicians, and this categorization plays a role in the inference process. Like the judgments studied by Kahneman and Tversky, category-based judgments occur frequently in everyday life, and seem to be based on heuristics rather than normative principles (Osherson, Smith, Wilkie, Lopez, & Shafir, 1990). Such judgments are the concern of this paper.

Distinctions

To structure our report, we need to introduce some terminology and distinctions. Note first that any probability question can be characterized as an argument, in which the known propositions are the *premises* of the argument and the proposition whose probability is in question is the *conclusion* of the argument. To illustrate, the preceding example amounts to the following argument (where the statement above the line is the premise, and the one below the line is the conclusion):

(1) A majority of surgeons oppose socialized medicine

$$\frac{\text{A majority of surgeons oppose socialized medicine}}{\text{A majority of internists oppose socialized medicine}}$$

In providing further examples, it is useful to switch to biological categories like birds and mammals, because there seems to be more consensus among people about the subsets of such categories than about the subsets of social categories. Two further examples of category-based arguments are:

(2) Robins have sesamoid bones
Ducks have sesamoid bones

All birds have sesamoid bones

(3) Robins have sesamoid bones
Ducks have sesamoid bones

Sparrows have sesamoid bones

In both cases, a subject might be asked to estimate the probability that the conclusion is true given that the premises are true.

Arguments like (2) are distinguished by the fact that the conclusion category, BIRD, properly includes the premise categories, ROBIN and DUCK; such arguments are said to be "general" (from here on in, we use capitals to indicate categories). In arguments like (3), in contrast, all categories are at the same hierarchical level; these arguments are said to be "specific". In this paper, we focus on specific arguments, though there will be some mention of general arguments.[1]

Another distinction concerns properties or predicates. *Having sesamoid bones* is a recognizable biological property, but not one that most people are familiar with, or can readily reason about. Such predicates are called "blank". They are to be distinguished from non-blank predicates like *can fly faster than 20 miles an hour*; we are familiar with, and can reason about, such a predicate. A rough test of whether a predicate is blank or not is whether it applies equally to all categories in a domain, or instead characterizes some categories better than others. *Having sesamoid bones*, for instance, seems no more likely of one bird species than another, whereas *can fly faster than 20 miles per hour* clearly characterizes some birds (e.g., hawks and eagles) better than others (e.g., chickens and ducks). This distinction between blank and non-blank predicates is a major concern of the present paper. To preview our results, we will show that: with blank predicates, judgments of probability are based mostly on similarity and typicality relations between premise and conclusion categories, just as they are in

[1] Specific arguments can also be distinguished by the fact that any natural category (e.g., BIRD) that properly includes any of the premise categories or the conclusion category properly includes the others as well. Arguments that are neither specific nor general are referred to as "mixed". For discussion of mixed arguments, see Osherson et al. (1990).

the representativeness heuristic; with non-blank predicates, however, probability judgments are based not only on similarity relations, but also on the plausibility of the premises and conclusion.

With this as background we can state our agenda. In the next section we consider category-based arguments with blank predicates. We will be brief here because much of the relevant research has appeared elsewhere (see Osherson et al., 1990; Osherson, Stern, Wilkie, Stob, & Smith, 1991; Smith, Lopez, & Osherson, 1992). In the third section we turn our attention to non-blank predicates. We present a model of how people reason about such predicates when judging the strength of arguments, along with some relevant data. Concluding remarks occupy the fourth and final section.

Category-based arguments with blank predicates

Factors that affect probability judgment

We are interested in factors that affect probability judgments about specific arguments. To uncover these factors, we presented 40 University of Michigan undergraduates with a series of 24 arguments, and asked them to estimate the probability of each conclusion on the assumption that the respective premises were true. Certain pairs of arguments offered contrasts that differed on only one factor, and these contrasts provide evidence for a number of phenomena. Four such phenomena are considered below.

One contrasting pair of arguments consisted of:

(4a) Tigers use serotonin as a neurotransmitter
Cougars use serotonin as a neurotransmitter [.86]

Bobcats use serotonin as a neurotransmitter

(4b) Tigers use serotonin as a neurotransmitter
Cougars use serotonin as a neurotransmitter [.39]

Giraffes use serotonin as a neurotransmitter

The numbers in brackets indicate the average conditional probability that subjects assigned to that particular argument (i.e., the probability they assigned to the conclusion being true, given that the premises were true). The two arguments in (4) differ with respect to the similarity of the premise categories to the conclusion category, this similarity being greater in (4a) than (4b); clearly,

Table 1. *Some phenomena involving specific arguments*

Phenomenon	Stronger argument	Weaker argument
1. Premise–conclusion similarity	TIGER, COUGAR/BOBCAT	TIGER, COUGAR/ GIRAFFE
2. Premise–conclusion asymmetry	LION/BAT	BAT/LION
3. Premise typicality (after Rips, 1975)	HORSE/GOAT	PIG/GOAT
4. Premise diversity	CHIMPANZEE, FOX/ POLAR BEAR	WOLF, FOX/ POLAR BEAR

subjects favored conclusions that were more similar to the premises. We refer to this effect as the "premise–conclusion similarity" phenomenon.

Table 1 lists three phenomena involving specific arguments that emerged from this study plus a fourth phenomenon that is due to Rips (1975). The first column of the table names the phenomenon. The second and third columns give the premise and conclusion categories used in the contrasting pair of arguments that define the phenomena. Column 2 lists arguments judged more probable (the "stronger" arguments), and Column 3 the less probable (or "weaker") arguments. The arguments are presented in the format "premise category . . . / conclusion category", with the blank predicate being suppressed. The difference between the stronger and weaker argument is statistically significant in most cases by a sign test.

After premise–conclusion similarity, the next phenomenon listed in Table 1 is asymmetry. It is defined only for single-premise arguments, and reveals that such arguments need not be symmetric. In particular, a single-premise argument will be judged more probable when the more typical category is in the premise than in the conclusion. For example, lions having a particular property makes it more probable that bats do, than vice versa. This asymmetry phenomenon is closely related to a phenomenon reported by Rips (1975), which we term the "typicality" phenomenon. It is listed as the third phenomenon in Table 1; it says that, other things being equal, arguments with more typical premise categories (e.g., HORSE) are judged more probable than those with less typical premise categories (e.g., PIG), even when the similarity between premise and conclusion categories is kept constant. The fourth phenomenon in the table is "premise diversity". This phenomenon shows that, other things being equal, the more diverse, or dissimilar, the premise categories, the more probable the argument is judged. Chimpanzees and foxes sharing a common property makes it more probable that polar bears have it, than does the fact that wolves and foxes share the same property. Note that more diverse premise categories may not be more typical, or more similar, to the conclusion category; for example, in the preceding example, the occurrence of chimpanzees increases diversity but not typicality or similarity to the conclusion. In what follows, we focus on the four phenomena of Table 1.

The similarity coverage model

To explain the preceding phenomena, among others, Osherson et al. (1990) advanced the *similarity coverage* model. This is a model of argument strength, where "strength" refers to the extent to which belief in an argument's premises causes the reasoner to believe in the argument's conclusion. When an argument's predicate is blank, its strength is captured by the judged probability of the conclusion given the premises, since prior belief in the conclusion plays no role. In such cases, a model of argument strength can also serve as a model of conditional probability judgment. Although the similarity coverage model applies both to general and to specific arguments of this kind, we focus on the latter in what follows.

The model

According to the model, the judged probability of an argument depends on two variables:

(i) The similarity of the premise categories to the conclusion category.
(ii) The extent to which the premise categories "cover" the lowest-level category that includes the premise and conclusion categories.

The first, or "similarity", term is straightforward; the only wrinkle is that when there are multiple premises, similarity is determined by a maximum rule. For argument (3), for example, the similarity term consists of the maximum similarity of robins or ducks on the one hand, to sparrows on the other. One piece of evidence for the maximum rule is that argument (3) does not change much in strength if the premise about ducks is removed. More generally, the maximum rule captures the intuition that, when judging the probability that a conclusion category has a particular property, we pay most of our attention to the most similar premise category.

The second, or "coverage", term of the model requires more unpacking. Note first that it presupposes that subjects judging a specific argument generate a more general category, namely, the lowest-level category that includes the premise and conclusion categories. We refer to this category as the "inclusive" category. For argument (4a), the inclusive category might be FELINE; for argument (4b), the inclusive category might be MAMMAL. The introduction of an inclusive category captures the intuition that when informed, for example, that tigers and cougars have a property, the subject considers the possibility that all felines have the property and, therefore, that bobcats do. The judgment of a specific argument thus includes the generation of a general argument. The strength of this general argument is evaluated by computing the extent to which its premises, for example, tigers and cougars, "cover" the inclusive category, for example, FELINE.

We now need to explicate the notion of "coverage". Intuitively, members of a general category cover that category to the extent that, on average, they are similar to other members. As an aid to intuition, Fig. 1 contains a two-dimensional representation – obtained by multidimensional scaling – of the similarities between various instances of the concept FRUIT (Tversky & Hutchinson, 1986). Similarity here is reflected by closeness in the space. If we restrict our attention to the coverage provided by single members, typical members like APPLE or PLUM cover the space better than atypical members like COCONUT or OLIVE; that is, the average metric distance of APPLE or PLUM to all other instances in Fig. 1 is less than that of COCONUT or OLIVE to all other instances in Fig. 1. The fact that a typical member provides relatively good coverage of a category gives us insight into why single-premise arguments with typical premise categories tend to be judged stronger than arguments with atypical premise categories, as revealed in the asymmetry and typicality phenomena. If we consider the coverage provided by multiple members of a category, however,

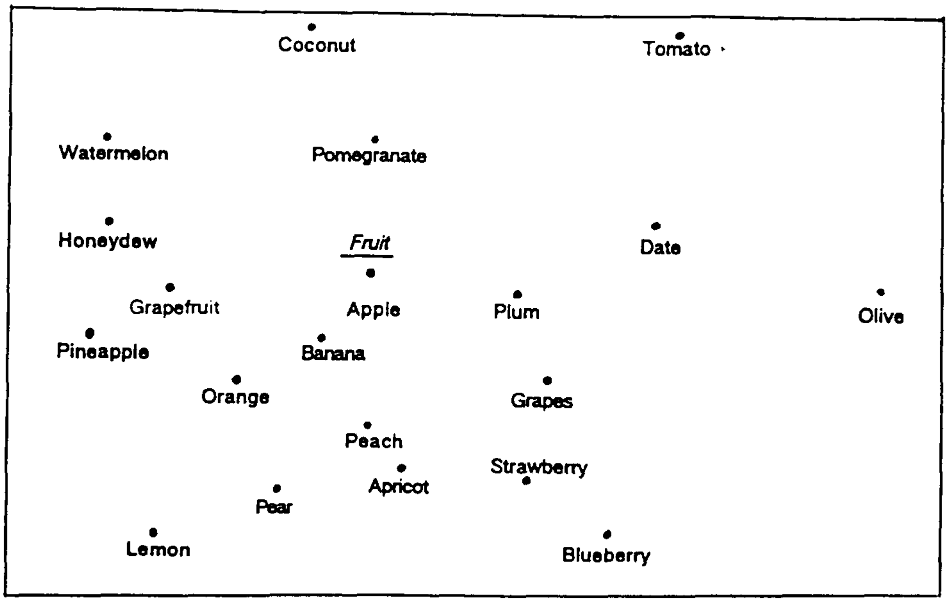

Figure 1. *A two-dimensional space for representing the similarity relations among 20 instances of fruit. From Tversky and Hutchinson (1986).*

more than typicality is involved. Intuitively, APPLE, PLUM, and ORANGE cover the space in Fig. 1 less well than do COCONUT, GRAPEFRUIT, and BLUEBERRY, even though the former are generally more typical than the latter. This difference in coverage arises because whatever category member is close to APPLE is also close to PLUM and ORANGE, so PLUM and ORANGE add little by way of coverage; in contrast, there are some members that are close to GRAPEFRUIT or BLUEBERRY but not to COCONUT, so GRAPEFRUIT and BLUEBERRY are adding coverage. This gives us insight into the diversity phenomenon.

Thus, a subset of a general category covers that category to the extent that, for any member of the latter you think of, at least one member of the former is similar to it. This statement leads naturally to an algebraic definition of coverage. Let $P_1 \ldots P_m/C$ be a general argument with premise categories CAT $(P_1) \ldots$ CAT(P_m) and conclusion category CAT(C). Furthermore, let $c_1 \ldots c_n$ be instances of CAT(C) that a person judging the argument considers (perhaps unconsciously). And let SIM(CAT(P_i), c_j) be the similarity between premise-category P_i and conclusion-category instance c_j. Then the coverage of an argument, which we will denote by COV(CAT(P_1) ... CAT(P_m); CAT(C)), is defined as the average of:

$$\begin{aligned}
&\mathrm{MAX}[\mathrm{SIM}(\mathrm{CAT}(P_1), c_1), \ldots, \mathrm{SIM}(\mathrm{CAT}(P_m), c_1)]\\
&\mathrm{MAX}[\mathrm{SIM}(\mathrm{CAT}(P_1), c_2), \ldots, \mathrm{SIM}(\mathrm{CAT}(P_m), c_2)]\\
&\cdot\\
&\cdot\\
&\cdot\\
&\mathrm{MAX}[\mathrm{SIM}(\mathrm{CAT}(P_1), c_n), \ldots, \mathrm{SIM}(\mathrm{CAT}(P_m), c_n)]
\end{aligned}$$

Coverage, then, is the average maximum similarity between (sampled) conclusion-category instances and premise categories. Returning to argument (4a), for which FELINE is presumably the inclusive category, the coverage term of the argument would be given by COV(TIGER, COUGAR; FELINE), and might include terms like:

MAX[SIM(TIGER, LEOPARD), SIM(COUGAR, LEOPARD)]; and
MAX[SIM(TIGER, HOUSECAT), SIM(COUGAR, HOUSECAT)]

Note that in the special case of single-premise arguments, such as TIGER/COUGAR, there is no maximum to consider. Hence, COV(CAT(P); CAT(C)) is simply the average similarity of the premise category to instances of the conclusion category. Coverage therefore reduces to the typicality of the premise category in the conclusion category (this fits with our geometric representation in Fig. 1).

Combining the similarity and coverage terms, our final model of argument strength is given by:

$$\begin{aligned}&(\alpha)\,\mathrm{MAX}[\mathrm{SIM}(\mathrm{CAT}(\mathrm{P}_1), \mathrm{CAT}(\mathrm{C})), \ldots, \mathrm{SIM}(\mathrm{CAT}(\mathrm{P}_m), \mathrm{CAT}(\mathrm{C}))]\\&+(1-\alpha)\,\mathrm{COV}[(\mathrm{CAT}(\mathrm{P}_1), \ldots, \mathrm{CAT}(\mathrm{P}_m); \text{INCLUSIVE CATEGORY}]\end{aligned}$$

The positive constant $\alpha(0 \leq \alpha \leq 1)$ indicates the weight given to the similarity term: the weight given to coverage is simply $1-\alpha$.[2]

Applications of the model

The similarity coverage model readily explains the four phenomena described earlier. The first phenomenon was premise–conclusion similarity, and we illustrated it by showing that an argument with premise categories TIGER and COUGAR and conclusion category BOBCAT is judged more probable than an argument with the same premise categories but conclusion category GIRAFFE. Assume that the inclusive category for the former argument is FELINE and that for the latter argument is MAMMAL. Then, according to the model, the strengths of the two arguments are given by:

(a) $(\alpha)\,\mathrm{MAX}[\mathrm{SIM}(\text{TIGER}, \text{BOBCAT}), \mathrm{SIM}(\text{COUGAR}, \text{BOBCAT})]$
$+(1-\alpha)\,\mathrm{COV}[\text{TIGER}, \text{COUGAR}; \text{FELINE}]$; and

(b) $(\alpha)\,\mathrm{MAX}[\mathrm{SIM}(\text{TIGER}, \text{GIRAFFE}), \mathrm{SIM}(\text{COUGAR}, \text{GIRAFFE})]$
$+(1-\alpha)\,\mathrm{COV}[\text{TIGER}, \text{COUGAR}; \text{MAMMAL}]$

The similarity term is clearly greater for (a) than (b). The coverage term is also greater for (a) than (b), because tigers and cougars are more similar on average to other cats than to other mammals. This explains why argument (4a) is judged the stronger.

Consider next the typicality phenomenon, which we illustrated by the argument HORSE/GOAT being judged more probable than the argument PIG/GOAT. Because Rips (1975) selected these items so that the similarity between premise and conclusion categories was held constant, we can focus on just the coverage term. Since MAMMAL is presumably the inclusive category for both arguments, for the model to predict the result of interest, the coverage of (HORSE; MAMMAL) must exceed that of (PIG; MAMMAL). Because coverage for single-premise arguments reduces to typicality, and because HORSE is in fact a more typical MAMMAL than is PIG, the result is accounted for. By a comparable line of reasoning, the asymmetry phenomenon is accounted for.

The last phenomenon to consider is that of premise diversity, which we

[2] In Osherson et al.'s (1990) analysis, coverage and α are defined for each particular individual. In the present treatment they are assumed to be the same for all individuals under consideration for ease of exposition.

illustrated by showing that CHIMPANZEE, FOX/POLAR BEAR is judged stronger than WOLF, FOX/POLAR BEAR. The items were selected so that the premise categories that differ between the arguments, CHIMPANZEE and WOLF, were roughly equally similar to the conclusion category. Hence, the similarity terms for the two arguments are roughly the same, and again we can focus on the coverage terms. Assuming that the inclusive category for both arguments is MAMMAL, coverage for the stronger or more probable argument might include terms such as:

MAX[SIM(CHIMPANZEE, SQUIRREL), SIM(FOX, SQUIRREL)]; and
MAX[SIM(CHIMPANZEE, MONKEY), SIM(FOX, MONKEY)]

Coverage for the weaker argument might include terms such as:

MAX[SIM(WOLF, SQUIRREL), SIM(FOX, SQUIRREL)]; and
MAX[SIM(WOLF, MONKEY), SIM(FOX, MONKEY)]

When the sampled conclusion instance is SQUIRREL, there should be little difference between the two arguments; that is, it is doubtful that WOLF is appreciably more similar to SQUIRREL than FOX is. However, when the sampled instance is MONKEY, the maximum similarity will be greater in the more diverse argument because CHIMPANZEE is more similar to MONKEY than WOLF is. More generally, only for the diverse argument will there be some conclusion instances that are covered by the second premise but not the first. Considerations like these account for the premise diversity phenomenon.

The preceding applications of the similarity coverage model comprise only a small part of the empirical support for the model. Osherson et al. (1990) present other phenomena that are predicted by the model. This same paper also presents the results of several experiments which show that the similarity coverage model provides a quantitative account of strength or probability judgments to specific and general arguments. Smith et al. (1992) provide additional experiments and quantitative tests of the model, and Osherson et al. (1991) show that a variant of the model quantitatively predicts strength or probability judgments on an individual subject basis.

Both the similarity and coverage terms of the model reflect similarity relations. This makes the model a close relative to the representativeness heuristic of Kahneman and Tversky (1973), which often reduces to similarity and typicality (Shafir et al., 1990). Things change substantially, however, when the predicates in our paradigm are accorded a more familiar content.

Category-based arguments with non-blank predicates

Counter-examples to the phenomena

Some of the phenomena that we have studied with blank predicates are extremely robust. The similarity and typicality phenomena, for example, have been obtained with preschool children (Lopez, Gunthil, Gelman, & Smith, 1992; see also Carey, 1985, and Gelman & Markman, 1987), and with a wide variety of categories, including artifacts, number categories, and social categories (Armstrong, 1991; Rothbart & Lewis, 1988; Sloman & Wisniewski, 1992). The robustness of the similarity phenomenon would seem to result from its close relation to the principle of stimulus generalization (similar stimuli occasion similar responses). It is thus of considerable interest that counter-examples to the phenomena can be generated by changing the predicates from blank ones to more familiar ones.

Consider the following pair of arguments (where "poodles" refers to "toy poodles"):

(5a) Dobermanns can bite through wire

German shepherds can bite through wire

(5b) Poodles can bite through wire

German shepherds can bite through wire

Note first that the predicate, *can bite through wire*, is non-blank. We are relatively familiar with its contents, and we can reason about it. Also, it clearly meets the criterion for non-blankness that we mentioned earlier: it applies differentially to various members of a domain (it applies more to Dobermanns than to poodles, for example). What is of particular importance is that, for most people, argument (5b) is stronger than argument (5a). The intuition behind this judgment seems to be that, "If even poodles can do it, surely German shepherds can". But this judgment violates the similarity phenomenon, since poodles seem less similar to German shepherds than do Dobermanns.

In a like manner, arguments (6a) and (6b) offer a counter-example to the typicality phenomenon:

(6a) Collies can bite through wire

German shepherds can bite through wire

(6b) Poodles can bite through wire

German shepherds can bite through wire

Informal judgments favor the second argument as the stronger one, even though collies are more typical dogs than poodles are. Arguments (7a) and (7b) provide a counter-example to the asymmetry phenomenon described earlier:

(7a) Collies can bite through wire

Poodles can bite through wire

(7b) Poodles can bite through wire

Collies can bite through wire

Again, informal judgments find the second argument as stronger ("If poodles can do it, likely collies can too"), even though the more typical category appears in the premise of argument (7a), not in that of (7b). In like manner, one could produce counter-examples to the premise diversity phenomenon.

What lies behind these counter-examples? As we will see, it is not that similarity relations cease to play a role when non-blank predicates are used. Rather, similarity is being overshadowed by another factor in these examples – the plausibility of the premise – where a less plausible premise, once accepted as true, seems to induce greater belief in the corresponding conclusion than does a more plausible premise. What we now need to do is to explicate the notion of plausibility.

The gap model

Basic ideas

The intuitions that lie behind our model for non-blank predicates are captured by the following hypothetical monologue of a subject judging arguments (5b) and (5a), respectively:

> "Hmm. Poodles can bite through wire. I thought that to bite through wire, an animal had to have powerful jaws. But poodles are kind of weak. I guess an animal doesn't have to be that powerful to bite through wire. Then a German shepherd is almost certainly powerful enough to do it."

> "Hmm. Dobermanns can bite through wire. That fits with what I thought, to

bite through wire an animal has to have powerful jaws. I'm not sure whether or not a German shepherd is powerful enough to do it."

The key ideas are:

(1) The non-blank predicate potentiates a subset of the premise category's attributes (e.g., powerfulness, size). From here on, these are the only relevant attributes. The non-blank predicate is associated with one or more of these attributes (e.g., powerfulness) and with values on them.

(2) The premise category (e.g., POODLE) is evaluated to see if its values on the relevant attributes are at least as great as those that characterize the predicate. The predicate's values are used as criteria that the category must pass.

(3) If the premise category's values are less than those of the predicate (e.g., POODLE's powerfulness level is less than the criterion set up by the predicate), the latter are scaled down; otherwise, the predicate's values are left unchanged (as in the Dobermann example). In this way, the plausibility of the premise plays a role.

(4) To the extent that the predicate's values are scaled down, the conclusion category's values are more likely to be at least as great as those of the predicate, and hence the conclusion is likely to be judged more probable. In this way, the premise indirectly affects the plausibility of the conclusion.

These processes differ from those underlying the similarity coverage model in that they require a decomposition of the predicate into its constituent attributes and values. In the similarity coverage model, a predicate is essentially treated as a whole (presumably because it is blank).

Probability of statements

We now embody the foregoing intuitive account into an explicit model that seems to offer the simplest possible realization of the key ideas. The model can be viewed as having two functions, which correspond to estimating the probability of an individual statement and estimating the probability of a conclusion of an argument given its premises. In what follows, we illustrate how the model works by relying on the poodle and Dobermann examples. (For a more extended discussion of the model, see Osherson, Smith, Myers, Shafir, & Stob, in press.)

Let us assume that the categories and predicate of interest have the attribute value structure given in Table 2. The values of attributes are assumed to correspond to real numbers (this amounts to beliefs being represented by real vectors in an appropriate attribute space). Note that the final column in the table gives the values of 9 and 9, which represents the powerfulness and size level required of an animal to be able to bite through wire. Presumably, these values,

Table 2. *A. Hypothetical attributes and values associated with three categories and one predicate figuring in arguments (5)–(7)*

	Poodles	Dobermanns	German shepherds	Can bite through wire
Powerfulness	3	7	6	9
Size	2	6	7	9

B. Hypothetical similarity ratings

Poodles, German shepherds = .3
Dobermanns, German shepherds = .6

unlike the others in the table, are not part of a pre-existing representation. Rather, by some knowledge-based processes, people are able to compute such attribute-value representations for predicates "on-the-fly". In essence, people may treat the predicate as an *ad hoc category* (ANIMALS THAT BITE THROUGH WIRE), and use whatever knowledge generation processes they typically use when constructing such makeshift categories (Barsalou, 1983).

Consider how the model determines the probability of a statement. For a statement to be probable, the values of the category should be at least as great as the corresponding values of the predicate. This idea may be quantified with the "cut-off" operator $\dot{-}$, defined over real numbers by:

$$x \dot{-} y = \mathrm{Max}\{0, x - y\}$$

(Thus $5 \dot{-} 3 = 2$ and $3 \dot{-} 5 = 0$.) Now suppose there are n relevant attributes. Letting C designate the vector for the category's values and P the vector for the predicate's values, the probability of a category-property statement is estimated to be:

$$\frac{1}{1 + \sum_{i=1}^{n} (\mathrm{P}_i \dot{-} \mathrm{C}_i)} \tag{1}$$

where P_i and C_i are the values at the ith coordinate (attribute) of the vectors P and C. To illustrate, according to Eq. (1) and Table 2, the probability that German shepherds can bite through wire is:

$$\frac{1}{1 + [(9 \dot{-} 6) + (9 \dot{-} 7)]} = 0.17 \tag{2}$$

Equation (1) will always yield a number in [0, 1]. Probability 1 is attained if $\mathrm{C}_i \geqslant \mathrm{P}_i$ for all attributes, that is, if the category's values satisfy all the criteria set by the predicate (the surplus of C_i over P_i plays no role in the calculation). Since the fundamental entity, $\mathrm{P}_i \dot{-} \mathrm{C}_i$, may be conceived as the "gap" between the predicate's and category's value on attribute i, the present theory is called the "gap model". These gaps are what lie behind the plausibility (or implausibility) of a statement.[3]

Conditional probability of statements

Now that we know how the gap model assigns a probability to an individual statement, we can consider how the model assigns a probability to a conclusion of an argument given its premises. Assigning such a conditional probability requires a number of steps. These include an initial modification of the argument's predicate due to the impact of the premises, and subsequent estimation of the conclusion's probability. The following example illustrates how these steps are implemented in the model.[4]

Argument (5b) may be abbreviated as:

$$\text{(5b)}\quad \frac{\text{(POODLE, WIRE)}}{\text{(GERMAN SHEPHERD, WIRE)}}$$

where each category is assumed to be represented by its values. In the first step, the subject assesses the impact of the premise by comparing the category's critical values to those of the predicate. On the attribute of powerfulness, the relevant gap is $\text{WIRE} \dot{-} \text{POODLE} = (9 \dot{-} 3)$ (see Table 2). This signifies that poodles do not have thc powerfulness needed to bite through wire. However, (POODLES, WIRE) is a premise, and hence assumed to be true. Therefore animals like poodles need not have a powerfulness of 9 to bite through wire. The subject is thus led to lower the powerfulness value for WIRE when evaluating the conclusion. In doing this, he or she considers not only the gap between WIRE and POODLE, but also the similarity between poodles and German shepherds. The similarity between the two kinds of dogs approximates the perceived relevance of the premise to the conclusion, and thus affects the extent to which the conclusion predicate is changed. Hence, the powerfulness value for WIRE is lowered by:

$$(\text{WIRE}_1 \dot{-} \text{POODLE}_1) \times \text{similarity}\,(\text{POODLE, GERMAN SHEPHERD}) \tag{3}$$

[3] A potential problem with Eq. (1) is that any addition of an attribute to the category and predicate tends to lower the probability of statements, since probability declines with the sum of the gaps. This aspect can be changed by taking an average of the gaps. To keep things simple, though, we will leave Eq. (1) as is.

[4] For an argument with a non-blank predicate, the judged probability of its conclusion given its premises is not a pure measure of the argument's strength. This is because the probability judgment may reflect one's prior belief in the argument's conclusion as well as the extent to which belief in the premises causes one to believe in the conclusion. For this reason, in both our examples and our experiments, we usually compare arguments that have the identical conclusion.

where the subscript 1 indicates that we are dealing with the first attribute. Hypothetical similarity values are provided in the bottom half of Table 2. Plugging the relevant numbers into Eq. (3) gives:

$$(9 \dot{-} 3) \times 0.3 = 1.8 \tag{4}$$

Thus the powerfulness value of WIRE is lowered by 1.8, so the modified value is $9 - 1.8$ or 7.2.

By a similar logic, the second attribute in Table 2, size, also gives rise to a gap, $\mathrm{WIRE}_2 \dot{-} \mathrm{POODLE}_2$. As a consequence, WIRE's size value is lowered by:

$$(9 \dot{-} 2) \times 0.3 = 2.1 \tag{5}$$

The modified value for WIRE's size is therefore $9 - 2.1$ or 6.9.

The premise (POODLE, WIRE) has thus modified the WIRE representation from its original values of (9, 9) to the new values of (7.2, 6.9). It remains only to consider the last step of the process, and calculate the probability of the conclusion. This step was explicated in our description of how one estimates the probability of individual statements. To implement this step, we need to insert into Eq. (1) the two gaps that involve GERMAN SHEPHERD and the modified WIRE – namely $7.2 \dot{-} 6$ and $6.9 \dot{-} 7$:

$$\frac{1}{1 + 1.2 + 0} = 0.45 \tag{6}$$

Observe that the latter probability exceeds the unconditional probability that German shepherds can bite through wire, calculated earlier to be 0.17 (see Eq. (2)). The difference is due to the impact of the premise (POODLE, WIRE), which changes the subject's interpretation of WIRE, bringing it into greater conformity with the values of POODLE.

It is instructive to go through comparable calculations for argument (5a), which may be abbreviated as:

(5a) (DOBERMANN, WIRE)

―――――――――――――――

(GERMAN SHEPHERD, WIRE)

In assessing the impact of the premise, the gap for powerfulness is $(9 \dot{-} 7)$ (see Table 2). Given that Dobermanns are 0.6 similar to German shepherds (see Table 2), the powerfulness value of WIRE is lowered by

$$(9 \dot{-} 7) \times 0.6 = 1.2 \tag{7}$$

This makes the modified powerfulness value of WIRE $9 - 1.2$, or 7.8. The gap for size is $(9 \dot{-} 6)$ (see Table 2). Hence the size value of WIRE lowered by:

$$(9 \dot{-} 6) \times 0.6 = 1.8 \tag{8}$$

making the size value of WIRE $9 - 1.8$, or 7.2. The values for the modified WIRE representation are therefore (7.8, 7.2), and the two gaps involving GERMAN SHEPHERD and the modified WIRE are consequently $7.8 \dot{-} 6$ and $7.2 \dot{-} 7$. Inserting these gaps into Eq. (1) gives the probability of the argument's conclusion as

$$\frac{1}{1 + 1.8 + 0.2} = 0.33 \tag{9}$$

Again the conditional probability exceeds the unconditional probability that German shepherds can bite through wire (0.17). The conditional probability obtained with DOBERMANN as premise category, 0.33, is less than that obtained with POODLE as premise category, 0.45. This fits our informal finding that argument (5a) is judged less probable than (5b).

This convergence of theory and data is merely demonstrational, since the critical parameters in Table 2 were generated for purposes of illustration. Indeed, small changes in the values of some entries in Table 2 would lead to the DOBERMANN argument being predicted to be stronger than the POODLE one (e.g., increasing the similarity between Dobermann and German shepherds could change the prediction). This implies that, when interpreting data with respect to the gap model, we cannot always expect the effect of premise gaps to overwhelm that of premise–conclusion similarity; that is, when premise gaps favor one argument but premise–conclusion similarity favors another, which argument emerges as stronger depends on the specific parameter values. Intuitively, gaps capture what has been learned from the premise, similarity reflects the relevance of this to the conclusion, and the final impact depends on the magnitudes of both factors.

The fact that the similarity between premise and conclusion plays a key role in the gap model provides an important link to the similarity coverage model. In both models, something about the premise predicate is generalized to the conclusion category to the extent the premise and conclusion categories are similar. However, in the similarity coverage model, the predicate – assumed to be blank – is left intact, whereas in the gap model the predicate's values are modified before it plays a role in the evaluation of the conclusion.[5]

[5] Another difference between the two models concerns how similarity between categories is computed. In applications of the similarity coverage model, we have assumed that each category is intrinsically associated with a large set of attributes, and that similarity between categories is computed over all these attributes (Osherson et al., 1991). In applications of the gap model, we have typically assumed that only attributes potentiated by the predicate matter, and hence that similarity between categories considers only two or three attributes (Osherson et al., in press). In essence, a non-blank predicate picks out certain *criteria* and similarity is computed only with respect to these criteria (Medin, Goldstone, & Gentner, 1993).

Multiple premises

The gap model can be extended to deal with arguments containing more than a single premise. The basic idea is embodied in a "maximization" principle. Specifically, when evaluating the gap for each attribute, every premise category is considered and the category with maximal impact is selected for use, where "impact" is defined as the product of gap and similarity. To illustrate, suppose argument (5a) were enriched to include a premise about collies in addition to that about Dobermanns. In calculating the gap on the powerfulness attribute, COLLIE will lead to a larger gap than DOBERMANN, assuming that collies are judged less powerful than Dobermanns; assuming further that collies and Dobermanns are roughly equally similar to German shepherds (the conclusion category), COLLIE would then have greater impact than DOBERMANN. Hence, by the maximization principle, only the gap due to COLLIE would be used to modify the powerfulness value of WIRE. The maximization principle can be stated more generally for arguments with any number of predicates. A similar proposal, couched in a connectionist architecture, has been made by Sloman (1993).[6]

This completes our description of the gap model. Its most important psychological claims are worth highlighting:

(1) The mental representations of categories and predicates can in part be summarized by attribute value structures (by real vectors in an appropriate attribute space).

(2) A category–predicate statement is judged probable to the extent that the values evoked by the predicate are contained in the category.

(3) An argument's premises increase the probability of its conclusion by lowering the values presumed necessary for possession of the predicate.

(4) The impact of a premise depends on (a) the gap between its predicate and category values, and (b) the similarity of its category to that of the conclusion.

(5) The impact of multiple premises is governed by the maximization principle.

Some qualitative results

To provide an empirical test of the gap model, we asked subjects to judge the probabilities of various one-premise arguments. The arguments were selected so

[6] The maximization principle introduces a kind of coverage notion into the gap model. Thus, when a second premise is added to an argument, it is unlikely to ever have substantially greater maximal impact than the initial premise if its category has similar values to the initial premise category. Hence, adding a more diverse premise category can lead to more of a modification in the predicate. Note that this notion of coverage differs from that in the similarity coverage model in that it concerns a space of features rather than a space of objects.

that there would be substantial variation in premise plausibility (which reflects premise gaps) and premise–conclusion similarity.[7]

Procedure

Five different categories of animals were used: house-cats, lions, camels, hippos, and elephants. House-cats and lions seem relatively similar to one another, as do hippos and elephants; all other couplings yield less similar pairs. (These intuitions were supported by subjects' similarity ratings.) Two different predicates were used: *have skins that are more resistant to penetration than most synthetic fibers* and *have a visual system that fully adapts to darkness in less than 5 minutes*. Each predicate was used with each category. We expected the first, or SKIN, predicate to result in a more plausible proposition when attributed to larger animals, such as hippos and elephants, than to smaller ones, such as house-cats; we expected the second, or VISUAL, predicate to result in a more plausible proposition when attributed to felines than to the other mammals. (These expectations were also supported by subjects' ratings.)

For each predicate, there are 20 possible one-premise arguments: each of the 5 categories could appear in the conclusion, with any of the 4 remaining categories appearing in the premise. The 20 arguments involving the VISUAL predicate were presented before the 20 involving the SKIN predicate. Within each of these 2 blocks, the 20 arguments appeared in random order.

On each trial, the subject was first presented the conclusion of the argument, and asked to rate on an 11-point scale how likely it was that the statement was true (where 0 indicated the statement was definitely not true and 10 indicated the statement was definitely true). Then the premise was presented, and subjects were instructed to consider it true. Subjects then reported how likely it was that the original claim (the conclusion) was true given the new information. After completing their probability ratings, subjects rated the similarity of all pairs of animals on an 11-point scale (where 0 indicated minimal similarity, and 10 maximum similarity). The subjects were 20 University of Michigan undergraduates who were paid for their participation.

Results

Table 3 gives the probability estimates for the SKIN predicate. These estimates were derived by first averaging over the 20 subjects, and then dividing each average by 10 to yield a number between 0 and 1. The first column of the table presents the data for cases in which a conclusion was presented alone. The next four columns indicate how the probabilities changed when each possible premise was added. To illustrate, consider the first row of Table 3, which contains the data for arguments in which house-cats served as the conclusion. The .40 in the first

[7] We are indebted to Kevin Biolsi for his assistance in all aspects of the following experiments.

Table 3. *Probability estimates for arguments with predicate,* Have skins that are more resistant to penetration than most synthetic fibers

Conclusion alone	With premise			
House-cats	Lions	Camels	Hippos	Elephants
.40	.59	.47	.41	.38
Lions	House-cats	Camels	Hippos	Elephants
.55	.80	.61	.53	.53
Camels	House-cats	Lions	Hippos	Elephants
.60	.77	.73	.60	.60
Hippos	House-cats	Lions	Camels	Elephants
.79	.93	.78	.82	.85
Elephants	House-cats	Lions	Camels	Hippos
.80	.80	.80	.86	.87

Pairwise similarity ratings

	House-cats	Lions	Camels	Hippos	Elephants
House-cats	–	8.10	2.85	1.55	1.45
Lions		–	5.45	3.20	3.95
Camels			–	3.00	3.75
Hippos				–	7.35
Elephants					–

column means that the probability attributed to the statement HOUSE-CAT SKIN when it appeared alone was .40. Adding the premise of LION SKIN boosted the probability of the HOUSE-CAT SKIN conclusion to .59; adding the premise CAMEL SKIN boosted the probability of the conclusion to .47; and so on. The bottom of the table contains the average pairwise similarity ratings for the five animals; these ratings will be needed in interpreting the probability estimates.[8]

Some points of interest can be gleaned from looking just at the first column of the top of the table – just at the probabilities attributed to the conclusions when they appeared alone. The SKIN property is non-blank, since the probability with which it is attributed to the five mammals ranges from .40 to .80. Because each of these statements also serves as a premise (in an argument with a different conclusion), there is also a substantial variation in premise plausibility. In particular, the premise involving house-cats is the least plausible, whereas the premises involving hippos and elephants are the most plausible. (Within each row, the premises are increasingly plausible as one moves from left to right.)

The more implausible a premise, the larger the premise gaps. The clearest evidence for the effect of gaps *per se* on judgments would come from cases where premise implausibility varies but premise–conclusion similarity remains relatively

[8] Ideally, we should have obtained two sets of similarity ratings for these animals, one in the context of the SKIN predicate and the other in the context of the VISUAL predicate (see footnote 5).

constant. The row corresponding to camels at the top of Table 3 comes the closest to offering such a case, as there is relatively little variation in similarity ratings involving camels (see the bottom of Table 3). The probabilities assigned to arguments with CAMEL SKIN as their conclusion in fact monotonically decrease as the premise becomes more plausible; this is clear-cut evidence for a gap mechanism. In two of the other rows of Table 3 – those which have house-cats or lions in their conclusion – premise implausibility is positively correlated with premise–conclusion similarity; that is, in these two rows, as one moves across the premises from left to right, the premise becomes both more plausible and less similar to the conclusion. In these cases the gap model unequivocally predicts that probabilities should decline in moving from left to right, and this prediction is supported in both rows of interest. In the remaining two rows – those corresponding to hippos and elephants – premise implausibility and premise–conclusion similarity point in opposite directions (as the former decreases, the latter increases). In such cases the gap model makes no clear prediction (since the critical commodity in the model is the product of gap and similarity). For the row corresponding to elephants, similarity seems to dominate, as the estimated probabilities consistently increase with premise–conclusion similarity. For the row corresponding to hippos, the largest estimated probability goes with the most implausible premise, but aside from that the estimated probabilities track similarity.

The preceding observations received statistical support in a stepwise regression analysis. The dependent measure was the change in the probability of a conclusion occasioned by the addition of a premise (this eliminated the influence that prior belief in a conclusion has on the conditional probability judgments). Premise–conclusion similarity entered the regression model first, and the only other factor in the final model was the interaction between similarity and plausibility. The coefficient for the similarity factor was .74 ($p < .01$), and that for the interaction factor was $-.09$ ($p < .05$); the multiple correlation was $R = .70$ ($F(3,\ 16) = 5.22$, $p = .01$). The form of the interaction was that the effect of implausibility was greater for more similar premise–conclusion pairs; exactly this kind of interaction is predicted by the gap model because implausibility (gap) is multiplied by similarity.

Table 4 presents the comparable data for the VISUAL predicate. There is substantial variation in the probability of the predicate being attributed to the various animals, from .42 to .79, which means there is substantial variation in premise plausibility. (Again, within each row, the premises increase in plausibility as one moves from left to right, but now the felines are associated with high plausibility whereas the hippos and elephants are associated with low plausibility.)

As before, first we consider the results for arguments that have camels in their conclusions, this being the closest we can come to a case where premise–conclusion similarity remains constant. Although there is not much variation in

Table 4. *Probability estimates for arguments with predicate,* Have a visual system that fully adapts to darkness in less than 5 minutes

Conclusion alone	With premise			
Hippos	Elephants	Camels	Lions	House-cats
.42	.63	.47	.42	.43
Elephants	Hippos	Camels	Lions	House-cats
.47	.69	.54	.52	.41
Camels	Hippos	Elephants	Lions	House-cats
.53	.55	.58	.56	.51
Lions	Hippos	Elephants	Camels	House-cats
.74	.66	.71	.76	.86
House-cats	Hippos	Elephants	Camels	Lions
.79	.74	.81	.75	.93

the judgments in the relevant row, there is at least some indication of a decline in probability estimates as the premise becomes more plausible. Moving to the first two rows of Table 4, those for hippos and elephants, we have cases in which both premise implausibility and similarity decline as one moves across the premises. Probability estimates systematically decline as one moves from left to right, as predicted by the gap model. In the last rows, which contain the data for lions and house-cats, plausibility and similarity point in opposite directions. In both of these cases, similarity seems to dominate. Thus, overall, premise–conclusion similarity seems to be playing a somewhat greater role in these data than in the data for the SKIN predicate considered previously.

A stepwise regression analysis again provides statistical support for our qualitative observations. Using increments in probability estimates as the dependent measure, the first factor to enter the model was premise–conclusion similarity, and the other factor in the final model was the interaction between similarity and plausibility. The coefficient for similarity was .52 ($p < .01$), that for the interaction was $-.04$ ($p < .01$), and the multiple correlation was .89 ($F(2, 17) = 32.92$, $p < .01$). Again, the form of the interaction was that predicted by the gap model, the effect of implausibility being greater when premise–conclusion similarity is greater. One last result is worth noting because it supports our observation that premise–conclusion similarity played a greater role with the VISUAL than the SKIN predicate. The simple correlation between similarity and probability estimates was .83 for the data obtained with the VISUAL predicate, versus .57 for the SKIN data. This difference in correlation is significant ($p <$ .05).

Why did the similarity factor dominate the plausibility factor more with the VISUAL predicate than the SKIN predicate? Two possible answers are worth considering. One possibility is that whether premise–conclusion similarity dominates implausibility (or vice versa) depends on the specific values of the relevant attributes and the specific values of the pairwise similarities (see p. 83). Thus, the particular values for our five animals on the relevant attributes of SKIN (e.g., size) may have been such as to produce a dominance of implausibility, whereas those on the attributes of VISUAL (e.g., nocturnal) may have produced a dominance of similarity. Alternatively, the SKIN and VISUAL predicates may differ in important qualitative respects. For example, the SKIN predicate seems to involve a *continuous* variation on an attribute like size – animals with very impenetrable skins tend to be large (elephant), those with medium impenetrable skin tend to be of medium size (lion), and those with relatively penetrable skin tend to be small (house-cat). In contrast, the VISUAL predicate may involve more of a *threshold* variation on some underlining attribute – either an animal has enough of this attribute to fully adapt to darkness in less than 5 minutes or it does not. If such a qualitative difference in predicates was the case, the gap model would likely provide a better account of the predicate associated with continuous variation because this is the kind of variation assumed in the model.

Some quantitative results

Having provided some qualitative support for the gap model, we now consider quantitative evidence. We are interested in the model's ability to make predictions about subjects' probability judgments for various arguments. This issue has been investigated in an experiment by Osherson, Smith, Myers, Shafir, and Stob (in press), and in what follows we describe some of their major findings.

Procedure

The materials included the 5 categories and 2 predicates used in the previous experiment, along with a new set of 5 categories and 2 predicates. The 5 new categories of animals were bears, beavers, squirrels, monkeys, and gorillas; the 2 new predicates were *have three distinct layers of fat tissue surrounding vital organs* (hereafter, abbreviated as FAT) and *have over 80% of their brain surface devoted to neocortex* (hereafter, NEOCORTEX). We will refer to the new material as "Set 2", and to the ones taken from Experiment 1 as "Set 1". Note that, as is the case with Set 1, Set 2 provides variations in similarity (e.g., bears are similar to beavers, and monkeys to gorillas), and in the applicability of the predicate to the category (e.g., FAT applies better to bears than to monkeys, whereas NEOCORTEX shows the reverse pattern).

More arguments were used in this study than the previous one. The arguments were constructed from the materials within a set in the same manner as in the previous study (i.e., a premise or conclusion of an argument consisted of a pairing of one of the categories with a predicate). For each predicate, there were : (i) 5 zero-premise arguments, that is, 5 cases where only a conclusion appeared; (ii) 20 one-premise arguments; (iii) 30 two-premise arguments; (iv) 20 three-premise arguments; and (v) 5 four-premise arguments. Hence, there were 80 arguments per predicate (which exhausts all the possibilities with 5 categories), for a total of 160 different arguments for each set.

Each subject made probability judgments about 160 arguments, presented in random order. Half the subjects worked with arguments constructed from Set 1, and the other half worked with arguments from Set 2. The phrasing of the probability question differed from that in the previous study. Now the entire argument was presented at once, and subjects were simply asked, "What is the probability that [the conclusion is true] given that [the premises are true]?" The "given that" clause did not appear for zero-premise arguments. The subjects were 30 University of Michigan undergraduates who were paid for their participation.

Results

To fit the gap model to an individual subject's data we need to know the attribute values of the relevant categories and predicates for that subject. These values were estimated by the following four-step procedure.

(1) The 160 arguments evaluated by a given subject were divided into two sets of 80 corresponding to the predicate appearing therein. Each set of 80 arguments was treated separately; this essentially divided each subject into two halves, and in what follows we refer to these 60 data sets (2 for each of our 30 subjects) as "half-subjects."

(2) The 80 arguments of a given half-subject were again partitioned into two sets. One set was used to fix the parameters of the gap model (as described below); these are the "fixing" arguments. The other set was used to test the predictions of the model once its parameters were fixed; these are the "testing" arguments. Different kinds of partitions were used, but we will focus on those in which either 20, 30, or 50 randomly selected arguments were used as the fixing arguments, and the remaining 60, 50, or 30 arguments served as the testing arguments.

(3) We assumed that each category and predicate could be represented by just two attributes. This means that the 5 categories and 1 predicate appearing in the 80 arguments of a half-subject are each associated with 2 values, for 12 values in all. These 12 values were the parameters to be determined. For each half-subject, a procedure was employed to find values of the 12 free parameters that maximize the gap model's fit to the fixing arguments. To illustrate, consider the partition in which the fixing set contained 20 fixing arguments. Choice of the 12 parameters caused the gap model to assign probabilities to each of the 20 fixing arguments. These predicted probabilities were then compared to those produced by a given subject for the 20 arguments, and the correlation between the predicted and

observed probabilities provided a measure of goodness to fit. The set of 12 parameters that maximized this correlation was retained for use in the next step.

(4) 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 were compared to the corresponding probabilities generated by the subject, with the correlation between predicted and observed probabilities providing a measure of goodness of fit of the model. Note that there are *no* free parameters in this last step.

Table 5 provides the results of our model fitting. The first column specifies how many fixing arguments were used, and the second gives the median correlation obtained in step (4) over all 60 half-subjects. Even with only 20 fixing arguments, the correlation between predicted and observed probability estimates is .72. The correlation rises with more fixing arguments. A question of some interest is, to what extent are these good fits dependent on similarity playing a role? To answer this question, we considered a variant of the gap model that contained no similarity factor, and fit this variant to the data using our four-step procedure. The results of these fits are in parentheses in Table 5. In all cases, the fit of the model to data is significantly poorer when similarity no longer plays a role. This is further evidence that similarity is an important element in modelling judgments about non-blank predicates, just as it was in modelling judgments about blank predicates.

An alternative procedure for fitting the model involves first obtaining subjects' estimates of values for the relevant attributes; then the model can be fit to data with no free parameters. We have done some preliminary work with this procedure. After making probability judgments as in the preceding experiment, subjects were presented with a list of 30 attributes that had been preselected to be appropriate to the categories and predicates of interest. They rated the extent to which each attribute characterizes each category or predicate. Using these estimated values to predict the subjects' probability judgments, we found correlations between predicted and obtained probability that generally range

Table 5. *Median correlations between predicted and observed probability estimates for different partitions of fixing and testing arguments (numbers in parentheses are for a variant of the gap model in which similarity plays no role)*

Number of arguments used for fixing	Median correlation
20	.72 (.60)
30	.77 (.70)
50	.84 (.77)

between .40 and .50. This is substantially lower than the correlations reported in Table 5. Higher correlations may be obtained with the present procedure by allowing the relevant attributes to vary with the arguments, rather than using a single list of attributes for all arguments (in which case, some attributes may not be relevant for many of the arguments). Additional studies are envisioned in which we employ more sensitive techniques for determining what attributes figure in subjects' judgments of particular arguments.

Concluding comments

Summary

Estimating the likelihood of a proposition on the basis of previous information can be thought of as assigning a probability to the conclusion of an argument based on its premises. We have explored two kinds of psychological factors that influence such probability assignment. First, in the context of blank predicates, we showed that argument strength depends on similarity relations between premise and conclusion categories. These relations were captured in the similarity coverage model, which explains a number of robust qualitative effects, including the premise–conclusion similarity and premise typicality phenomena. Counterexamples to these phenomena, however, were shown to arise in the context of predicates that are non-blank. Unlike the case with blank predicates, in which little is known about the predicate, subjects presented with non-blank predicates are likely to reason about the nature of the predicate and its plausibility *vis-à-vis* the category. Indeed, argument strength with non-blank predicates was shown to be a function of the premises' and conclusion's plausibility, in addition to category similarity. As reflected in the gap model, the perceived implausibility of premises captures how much has been learned from the fact that they are supposedly true, and the similarity between premise and conclusion categories reflects the relevance of what has been learned to the argument's conclusion.

In moving from blank to non-blank predicates, the reasoner's reliance on similarity decreases. Indeed, for some non-blank predicates, similarity seems to play no role at all. To illustrate, consider the following arguments:

(8a) A rhino can break this kind of scale

An elephant can break this kind of scale

(8b) A refrigerator can break this kind of scale

An elephant can break this kind of scale

Both arguments seem equally strong despite the fact that premise–conclusion similarity is far greater in the first than the second case (presumably premise plausibility is comparable in the two arguments). In these cases, there is little or no uncertainty about which attributes are critical (i.e., weight seems to underly the predicate), and consequently there seems to be no need to consider the similarity between premise and conclusion categories (Douglas Medin, personal communication). More generally, it appears that: a high degree of confidence about the critical attribute, as in the preceding example, leads to an almost total reliance on premise plausibility; a lack of confidence about the critical attributes, as with blank predicates, leads to an almost exclusive reliance on similarity; and a limited degree of confidence about the critical attributes, as with most of the non-blank predicates considered in this paper, leads to a reliance on both plausibility and similarity. In this manner, we can see the connection between the gap model and the similarity coverage model.

Possible extensions

In what follows, we briefly note three possible extensions of our analysis of non-blank predicates.

Thus far, we have assumed that similarities are never negative. Some category pairs, however, may be perceived as more different than they are similar, which suggests they may be perceived as negatively similar (this suggestion is compatible with Tversky's, 1977, contrast model of similarity). Negative similarities seem especially possible when the predicates are non-blank. A non-blank predicate may be associated with only one or two attributes, and if two categories have values at the extremes of these attributes they may be perceived as contraries. To the extent that similarity is assumed to reflect the relevance of premise to conclusion, it may be that a similarity score around zero entails little or no relevance, whereas high positive and high negative similarities both reflect great relevance, but of opposite sorts. Thus, assuming we believe that mice and squirrels have similar eating preferences whereas mice and lions have very different preferences, we may think that the fact that mice love onions increases the chances that squirrels love them, but also increases the chances that lions do *not*. The introduction of negative similarities to the gap model has a number of interesting implications that we have only begun to explore.

A second extension of the gap model stems from its prediction that the addition of a premise can *never* lower the probability of a conclusion. That is, given the model's particular cut-off operator and its notion of critical values, an additional premise can only increase the strength of a conclusion (when non-negative gaps occur and when similarity is relatively high), or leave it unchanged (when the gaps are negligible or when similarity is relatively low). However, as demonstrated by Osherson et al. (1990), premises may sometimes lower conclusion probability when blank predicates are used, and this seems all the more likely with non-blank predicates. One way to remedy this potential problem is to introduce into the gap model coverage principles like those used in the similarity coverage model, since an additional premise may change the inclusive category (currently, the gap model makes no use of category structure). Another way to account for decreasing belief with increasing premises is to change the particular fashion in which attribute values are compared (i.e., replace the cut-off operator); and still another possible change is the introduction of negative similarities, as described above. If any of these changes were implemented, the model would be able to predict an occasional lowering of belief in an argument's conclusion with additional premises.

A third elaboration of our model stems from an extension of our paradigm. Up to this point we have used arguments in which the categories vary but the predicate remains fixed. It is equally possible to entertain arguments in which the category is fixed but the predicates vary, and strength variations in such arguments may also be captured by the gap model. Consider, for example, arguments (9a) and (9b):

(9a) Geese can attain flight speeds of 110 miles per hour

Geese can develop a "lift" in excess of twice their body weight by agitating their wings

(9b) Geese can develop a "lift" in excess of twice their body weight by agitating their wings

---

Geese can attain flight speeds of 110 miles per hour

Intuitively, (9a) seems stronger than (9b). This fits with the key ideas behind the gap model; the proposition about flight speeds seems more implausible than that about "lift", and an argument is judged stronger when the more implausible proposition is in the premise rather than the conclusion. However, whereas the fixed-predicate type of argument naturally leads us to a gap-induced modification of the shared predicate (which is then transferred to the conclusion), the current, fixed-category argument seems likely to involve modification of the shared category. That is, while in our previous discussion people were assumed to adjust a predicate's critical values in light of the categories that are seen to possess the predicate, the envisioned extension of the gap model would incorporate changes in the values of the category in light of the predicates that it is said to possess.

Finally, arguments that involve both varied categories and predicates would require a model that incorporates both kinds of adjustment processes.

A final comment

Reasoning and decision making require the constant estimation and readjustment of the probabilities of propositions. Our subjective probability estimates, based on what we have come to know and believe so far, affect the ways in which we reason about future outcomes, and the decisions that we make. The discovery of facts previously considered implausible lead us to adjust our views in more substantial ways than facts that we do not find surprising. The ways in which we adjust our views depend largely on our categorization schemes, and on the similarities that we perceive between the categories under consideration. Further work in the cognitive and decision sciences should help clarify the ways in which plausibility and similarity affect inductive inference and judgments of probability.

References

Armstrong, S.L. (1991). Category and property specificity in category based induction. Poster presented at Psychonomic Society, San Francisco.

Barsalou, L. (1983). Ad hoc categories. *Memory & Cognition*, *11*, 211–227.

Carey, S. (1985). *Conceptual change in childhood.* Cambridge, MA: MIT Press.

Gelman, S.A. & Markman, E. (1987). Young children's induction from natural kinds: The role of categories and appearances. *Child Development*, *58*, 1532–1541.

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

Lopez, A., Gunthil, G., Gelman, S.A., & Smith, E.E. (1992). The development of category based induction. *Child Development*, *63*, 1070–1090.

Medin, D.L., Goldstone, K., & Gentner, D. (1993). Respects for similarity. *Psychological Review*, *100*, 254–278.

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

Osherson, D.N., Smith, E.E., Myers, T.S., Shafir, E.B., & Stob, M. (in press). Extrapolating human probability judgment. *Theory and Decision.*

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

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

Rothbart, M., & Lewis, P. (1988). Inferring category attributes from exemplar attributes. *Journal of Personality and Social Psychology*, *55*, 861–872.

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

Sloman, S.A. (1993). Feature based induction. *Cognitive Psychology*, *25*, 231–280.

Sloman, S.A., & Wisniewski, E. (1992). Extending the domain of a feature-based model of property induction. *Proceeding of the 13th Annual Meeting of the Cognitive Science Society*, Bloomington, IN.

Smith, E.E., Lopez, A., & Osherson, D.N. (1992). Category membership, similarity, and naive induction. In A. Healy, S.M. Kosslyn, & R. Shiffrin (Eds.), *From learning processes to cognitive processes: Essays in honor of W.K. Estes* (Vol. 2, pp. 181–206). Hillsdale, NJ: Erlbaum.

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

Tversky, A., & Hutchinson, W. (1986). Nearest neighbor analysis of psychological spaces. *Psychological Review*, *93*, 3–22.

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