COGNITIVE PSYCHOLOGY 24, 341–380 (1992)

# Categorization Using Chains of Examples

EVAN HEIT

*Stanford University*

People can infer unknown features of a stimulus by retrieving memories of similar examples. It is proposed that we can reason from *chains* of examples. For example, stimulus *A* may remind us of *B*, which reminds us of *C*. Information about *C* may then affect reasoning about *A*. A mathematical model for categorization (extended from the context model of Medin & Schaffer, 1978), using multiple-step chains of reasoning and memory for examples, is presented. In five experiments, subjects memorized feature descriptions of fictional people, then made predictions from incomplete descriptions. Various predictions could be made using one-, two-, or three-step chains of reasoning. These experiments varied in terms of stimulus structure, complexity of test questions, and response method (probability estimate or forced choice). The multiple-step context model, with the assumption that people performed one- and two-step chains of inference, successfully accounted for the results of all five experiments. 

## INTRODUCTION

### *Inferring Any Feature of a Stimulus*

Psychological research on categories has had a limitation. While people can make many inferences about a given stimulus, experiments have typically investigated inferences only about one specific aspect of the stimulus: the designated category label. In most artificial category learning experiments, subjects are only tested by having them provide the label for an unlabeled stimulus.

When we encounter some animal, we might try to infer the category label for its species. But when we encounter a large, growling creature in the dark, we might be more interested in inferring other properties of this creature, such as whether it will attack and how fast it can run, rather than naming it. From a statistical viewpoint, there is no difference between

This research was supported by a National Science Foundation Graduate Fellowship and a dissertation grant from Stanford University. This paper is based on a dissertation submitted to Stanford. Experiments 1 and 2 were presented at the 1990 Meeting of the Psychonomic Society in New Orleans. I am grateful to G. Bower, W. Ahn, J. R. Anderson, J. Greeno, P. Johnson-Laird, E. Markman, D. Medin, M. Pavel, L. Rips, R. Shepard, E. E. Smith, and E. Wisniewski for comments on earlier drafts. Address correspondence to Evan Heit, Department of Psychology, University of Michigan, 330 Packard Road, Ann Arbor, MI 48104, or Evan.Heit@um.cc.umich.edu.



0010-0285/92 $9.00

predicting the category label of an object and predicting the value of some other feature of the object, although some features may be more important to predict than others (Anderson, 1991; Billman & Heit, 1988).[1] Category learning models are generally used to predict only category labels (e.g., Gluck & Bower, 1988; Medin & Schaffer, 1978; Nosofsky, 1984; Posner & Keele, 1968). However, some more complex learning models have been proposed to account for inferences between any features or category labels (e.g., Anderson, 1991; Billman & Heit, 1988; Holland, Holyoak, Nisbett, & Thagard, 1986; auto-associators in Rumelhart & Zipser, 1985).

I do not mean to claim that there is nothing special about category labels and names. Particularly for natural categories, there are differences between learning what kind something is, e.g., it is a grizzly bear, and simply learning some property of it, e.g., it is big (cf. Markman, 1989). However, even for natural categories, people can still infer other features besides the category label. And in artificial category learning experiments, there is little justification for treating what the experimenter designates as the category label (e.g., "Disease A") differently from the other stimulus features (e.g., "watery eyes").

*Categorization from Examples*

How do we make these inferences? Most theories of categorization may be described as instance models or abstraction models. Roughly speaking, instance models (e.g., Hintzman, 1986; Medin & Schaffer, 1978; Nosofsky, 1984) represent categories in terms of individual category exemplars, and abstraction models (e.g., Fried & Holyoak, 1984; Posner & Keele, 1968) represent categories in terms of summary information rather than instances. According to several researchers, much of our reasoning is based on memory for specific examples.

Brooks (1978) argued that often, when abstract knowledge is not directly available, people may learn categories in terms of instances instead of trying to form abstractions. For example, it might be desirable to directly store members of categories that would be hard to describe abstractly, and during conditions of uncertainty or limited processing resources. Markman (1989) reviewed evidence that, in learning the categories of their first language, children may sometimes rely on instance representation. Young children without the reasoning ability to test complex hypotheses about word meanings might instead keep examples in memory and make category inferences on the basis of similarity to these examples.

Ross, Perkins, and Tenpenny (1990) have also studied reminding-based

[1] By feature, I mean an aspect of the mental representation of an object. Thus a category label may be considered as a special type of feature.

categorization. When someone encounters a new test example, it leads to retrievals of similar examples, which guide the current classification. Retrieving these particular instances also affects subsequent classifications. Using trait descriptions of fictional persons as stimuli, Ross et al. (1990) found evidence that categorization in their experiments was based on distinct memories for examples.

In the field of artificial intelligence, the technique of *case-based reasoning* has been successful for solving complex tasks using memory for examples (Riesbeck & Schank, 1989). This technique involves making an inference about a current case by retrieving similar cases from memory. For example, the computer program MEDIATOR solves international conflicts by retrieving case studies of similar past conflicts and inferring what solutions should be attempted for the present case (Kolodner & Simpson, 1989).

*Reasoning from Chains of Examples*

We can use memories for examples to support powerful kinds of reasoning. Imagine that you are trying to determine whether a person named Clyde is pro-life or pro-choice on the subject of abortion. Clyde may seem similar to Fred and Henry, because all three share certain characteristics, including belonging to the same social club. If Fred and Henry are both pro-life, then you might infer that Clyde also shares this view. This process is standard instance-based categorization, and will be referred to as a *one-step inference*. However, what if Fred's and Henry's views on abortion are unknown? Perhaps Fred and Henry seem similar to George. If George is pro-choice, then you might infer that Clyde too is pro-choice. This categorization is based on a two-step chain of sequential retrieval, from Clyde to Fred and Henry, and from Fred and Henry to George. Clyde and George might not have seemed similar to each other before this operation. This kind of operation will be referred to as a *two-step inference*. While reasoning from chains of examples may sometimes lead to incorrect inferences, it might frequently prove valuable and even necessary. It is likely that our memories for particular things contain a lot of missing information, so that relations between objects are not obvious simply from direct feature overlap. The only way to use $A$ to infer that $B$ has property $P$ may be by way of a mediated chain of remindings, from $B$ to $C$ and from $C$ to $A$.

Other researchers have argued for the plausibility and usefulness of reasoning by chains. Osherson, Smith, Wilkie, López, and Shafir (1990) told subjects that various animals, $A_i$, had some property $P$, then the subjects had to assess their confidence that another kind of animal, $B$, had that property $P$. Osherson et al. (1990) found that their subjects were sensitive to both the direct similarity between the $A$ and $B$ items, and to

the similarity of the $A$ items to other animals, $C_i$, that are in the same category as, and thus similar to, $B$. People might infer that $B$ has property $P$ even if $A$ and $B$ are not similar to each other, because both $A$ and $B$ resemble some animals $C_i$.

Case-based reasoning techniques may employ multiple-step chains of reasoning (Riesbeck & Schank, 1989). If the computer program is directed to infer whether some example $B$ has property $P$, it can retrieve case $C$ on the basis of similarity between $B$ and $C$. If $C$ does not provide information about property $P$, then $C$ may be used to retrieve case $A$, which has information about property $P$. This technique could perform the Clyde task above.

Finally, multiple-step reasoning could be used to perform another sort of categorical inference: *property inheritance*. Item $B$ might resemble some items $C$, for which information about property $P$ is unknown. However, the $C$ items may be in the same category as, and similar to, some item $A$ that has $P$. Thus it may be inferred that $B$ belongs in the same category as $C$ and $A$, and thus that $B$ has property $P$. It would be valuable to extend our conception of instance models to allow property inheritance, because this important capability has not yet been demonstrated for these models of categorization (Kahneman & Miller, 1986, p. 139).

### *Overview*

In the next section, I present a mathematical model of reasoning from examples that is an extension of Medin and Schaffer's (1978) context model of categorization. This extended model has two novel characteristics: it allows inferences about any stimulus feature and it uses multiple-step chains of similarity. Five category learning experiments evaluated the novel characteristics of this model as descriptions of human reasoning. While the first three experiments looked at one- and two-step chains of reasoning, the fourth and fifth experiments also examined evidence of people performing longer, three-step chains of inference. The fourth experiment also examined how people's reasoning is influenced by the sample size of instances available.

## A MATHEMATICAL MODEL OF REASONING FROM EXAMPLES

### *The Context Model*

The context model of categorization (Medin & Schaffer, 1978) assumes that instances, composed of binary features, are stored in memory along with their category labels. For an experiment involving four binary features and two categories $A$ and $B$, a sample of relevant memory traces might be represented schematically as $(-1\ -1\ -1\ 1\ A)$, $(-1\ -1\ 1\ -1\ A)$, $(1\ -1\ -1\ -1\ A)$, $(1\ 1\ -1\ 1\ B)$, and $(1\ -1\ 1\ 1\ B)$, where the $-1$s and 1s

stand for the values of particular features. (The designations of 1 and $-1$ are arbitrary, and could be replaced by 0 and 1, or by another pair of symbols. They are chosen to reflect that each feature can take on one of two *opposite* values, rather than showing the presence or absence of a specific feature.) To classify a new instance, $x$, as a member of category $A$ or $B$, the subject compares $x$ to all of the $A$ and $B$ instances in memory. The similarity, $sim(x,y)$, between any two instances $x$ and $y$, which are each composed of $N$ binary features, is defined in Eq. (1) as

$$sim(x,y) = \prod_{i=1}^{N} s_i, \tag{1}$$

where $s_i = w_i$, $0 \leq w_i \leq 1$, if $x_i \neq y_i$, and $s_i = 1$, if $x_i = y_i$. The symbols $x_i$ and $y_i$ refer to the $i$th features of instances $x$ and $y$, respectively. The $w_i$ parameters are attentional weights on each feature, with smaller values for features that get more attention. If $x$ and $y$ completely match each other, then $sim(x,y) = 1$; $sim(x,y)$ will decrease as $x$ and $y$ fail to match on more features. The $w_i$ are estimated parameters; they are not predicted by the context model.

The probability of judging $x$ to be a member of category $A$ is a function of $x$'s total similarity to category $A$ members relative to category $B$ members, as shown in Eq. (2)

$$p(A/x) = \frac{\sum_{a \in A} sim(x,a)}{\sum_{a \in A} sim(x,a) + \sum_{b \in B} sim(x,b)}. \tag{2}$$

In this equation the probability of classifying $x$ as an $A$ increases as $x$ is more similar to members of category $A$ and decreases as $x$ becomes more similar to members of category $B$. This quantity may also be interpreted as a single subject's estimate of the likelihood that instance $x$ belongs to category $A$.

This form of the context model compares a new instance to *every* memorized instance. However, Medin and Schaffer (1978, p. 211) noted that the context model could be reformulated so that each subject is presumed to make a smaller number of comparisons (e.g., just one, randomly selected) for each judgment, but this alternate conception of their theory can produce the same results.

*Inferring any Feature of a Stimulus*

One limitation of the context model, as presented, is that it assumes that subjects only predict category labels. This model cannot predict the

value of one feature from another, or the value of a feature given the category label. Fortunately, it is easy to generalize the context model so that any feature $j$ may be treated as a category label and thus predicted. Let $U$ be the set of all relevant instance traces in memory, e.g., the stimuli in an experiment. Let $x$ be a test instance, missing the value of the $j$th feature. The *sim* function in Eq. (1) is then used to compute similarity across all features except $j$. Let $A$ be the subset of $U$ consisting of instances with feature $j$ having the value 1, and let $B$ be the subset of $U$ consisting of instances with feature $j$ having the value $-1$. With these identifications, Eq. (2) may be used to predict the probability of instance $x$ having 1 as the value of feature $j$. Another instance model, Minerva 2 (Hintzman, 1986), can also treat any feature as a category label. However, until now, neither Minerva 2 nor the context model have been applied to experimental situations in which people predict different features of a stimulus.

While it is simple to alter the context model to predict any feature, it would be more difficult to make this alteration for other models that learn specific relations between cues and responses. The network model of Gluck and Bower (1988), the category density model of Fried and Holyoak (1984), the prototype model of Rosch, Simpson, and Miller (1976), as well as multiple-cue probability learning models (Klayman, 1988), all learn unidirectional relations between particular features designated as input cues and particular features designated as output categories. Thus these models cannot explain the results of the present experiments.

*Reasoning from Chains of Examples*

The context model is restricted to inferences based on direct similarity between a test stimulus $x$ and examples in memory. Yet people may reason using a chain of sequentially retrieved examples. A subject might use memory for example $y$ to make an inference about example $x$, if both $x$ and $y$ are similar to some example $z$, even though $x$ and $y$ are not directly similar to each other. The *multiple-step context model* uses a new similarity function to describe this possible form of reasoning. Equation (2) above must be replaced by Eq. (3),

$$p(A/x) = \frac{\sum_{a \in A} multisim_k(x,a)}{\sum_{a \in A} multisim_k(x,a) + \sum_{b \in B} multisim_k(x,b)}. \quad (3)$$

The *multisim* function is defined in Eqs. (4 and 5):

$$multisim_k(x,y) = sim(x,y), \quad \text{if } k = 1 \quad (4)$$

$$multisim_k(x,y) = \sum_{z \in U} multisim_{k-1}(x,z)sim(z,y), \qquad \text{if } k > 1. \tag{5}$$

The $k$ parameter of the *multisim* function designates how many steps of similarity are used to relate $x$ and $y$. In the base case of $k = 1$, the function reduces to the standard *sim* function defined in Eq. (1), where only direct similarity between the two examples is considered. To understand this function, consider the case of $k = 2$ in Eq. (6),

$$multisim_2(x,y) = \sum_{z \in U} sim(x,z)sim(z,y). \tag{6}$$

The $multisim_2$ function can be thought of as "activation" spreading from the test example $x$ to all traces $z$ in memory, in proportion to the direct similarity between $x$ and $z$. Then activation is likewise spread from the $z$s to the trace $y$, in proportion to the similarity between $z$ and $y$.[2] The $multisim_2$ function allows the context model to perform two-step chains of inferences. It is assumed, as a simplification, that $U$ contains only stimuli learned in the experiment, because extraexperimental stimuli are likely dissimilar to $x$ and $y$, and would have little impact on reasoning.

Equation (5) may also take a $k$ parameter of greater than 2, to perform chains of reasoning of more than two steps. The recursive nature of Eq. (5) allows us to think of long chains as combinations of shorter chains. For example, the three-step similarity between $x$ and $y$ is equivalent to the sum, over all traces $z$ in memory, of the product of the one-step similarity between $x$ and $z$ and the two-step similarity between $z$ and $y$. Then, the two-step similarity between $z$ and $y$ can be computed, using the same recursive function, as the sum of products of pairs of one-step similarities.

The multiple-step context model, and the *multisim* function, have several notable properties. First, $multisim_k(x,y)$ is sensitive not only to similarity between $x$ and $y$ based on $k$-step chains, but also to similarity between $x$ and $y$ due to shorter chains, including direct similarity between the two examples. In the simple case of $k = 2$, the value of the function

[2] The *multisim* function is conceptually similar to the intertrace resonance process that has been added to Hintzman's Minerva 2 model. In this process, a stimulus probe causes an echo by activating all traces in memory, in proportion to the similarity between the probe and the trace. Then, in subsequent iterations, memory traces activate each other, according to their past activation and their similarity to each other. In a manner like the multisim$_2$ function, a stimulus probe may strongly activate some memory trace on the second echo, even if the trace was not activated strongly on the first echo. Minerva 2 with intertrace resonance has not been applied to multiple-step inference tasks, but it is likely that it would behave in a similar way to the multiple-step context model. However, Minerva 2 allows only qualitative accounts of experimental data, while the multiple-step context model makes precise numerical predictions.

is the sum of $sim(x,z)sim(z,y)$, when $z$ takes the values of all relevant examples in memory. This sum includes the term $sim(x,y)sim(y,y)$. By Eq. (1), $sim(y,y) = 1$; so the sum includes the term $sim(x,y)$. Thus the one-step similarity between $x$ and $y$ is incorporated in the two-step similarity. However, this model does not include a mechanism for "stopping"; whatever the value of $k$, every response will incorporate judgments of $k$ steps.

Second, the use of multiplication in Eq. (5) acts like an "and" function. It is not enough that $x$ and $y$ be closely similar to other examples $z$, but they must be closely similar to the same examples $z$. Equation (5) heavily weights chains composed solely of high similarities.

Finally, just as the context model need not be interpreted as a claim that each subject retrieves every category member from memory to make a decision, the *multisim* function need not be interpreted as a claim that each subject examines all possible $z$ traces along chains from $x$ to $y$. Instead, if an individual subject randomly samples a fixed number of $z$ from $U$, then the *multisim* function's output will be proportional to the expected value of a single subject's computation.

### *Illustration of the Multiple-Step Context Model*

In these experiments, subjects memorize a set of binary feature descriptions of people. Then, on each test item, they judge the likelihood of some feature value, given one or two other feature values. Table 1 illustrates a simple version of this task and how it would be modeled. Here the subject has memorized five person descriptions, each with information

TABLE 1
Illustration of Multisim Function

| Name<br>Memory trace: | $Multisim_1$<br>(Probe, trace) | $Multisim_2$<br>(Probe, trace) |
|---|---|---|
| Holly<br>1 1 ? | $s$ | $s(1) + s(1) + s^2(s^2) + s^2(s) + s^2(s^2) = 2s + s^3 + 2s^4$ |
| Judy<br>1 1 ? | $s$ | $s(1) + s(1) + s^2(s^2) + s^2(s) + s^2(s^2) = 2s + s^3 + 2s^4$ |
| Karen<br>−1 −1 ? | $s^2$ | $s(s^2) + s(s^2) + s^2(1) + s^2(s^2) + s^2(s) = s^2 + 3s^3 + s^4$ |
| Monica<br>? 1 1 | $s^2$ | $s(s) + s(s) + s^2(s^2) + s^2(1) + s^2(s^2) = 3s^2 + 2s^4$ |
| Patricia<br>? −1 −1 | $s^2$ | $s(s^2) + s(s^2) + s^2(s) + s^2(s^2) + s^2(1) = s^2 + 3s^3 + s^4$ |

*Note.* The symbol '?' stands for a missing value. Memory is probed with (1 ? ?), and the value of the third feature is predicted. It is assumed that each feature has the same attention weight, $s$.

about two of three binary features. Note that Holly and Judy have exactly the same description. Nosofsky (1988) showed that the context model performs better by assuming that multiple presentations of the same description each lead to a separate memory trace. Thus I will likewise assume distinct memory traces for Holly and Judy.

In this illustration, the subject judges the probability that the value of the third feature will be 1, given that the first feature has the value 1. The critical stimuli for making this judgment are Monica and Patricia, because only they have information about the third feature. Table 1 shows the one- and two-step similarities between the probe, (1 ? ?) and the various memory traces. The *sim* function was used to compute similarity based only on the first two features, since the third feature was to be predicted. It is assumed here that each feature has the attentional weight of $s$. The '?' symbol stands for a missing value, and a "?" was considered a mismatch with any other feature value, including another "?" Thus two features matched only if they were both "1" or both "$-1$."

The critical memory traces, Monica and Patricia, have the same one-step similarity to the probe, $multisim_1$(probe, Karen) = $multisim_1$(probe, Patricia) = $s^2$ [see Eqs. (4 and 1)]. Thus the one-step context model would predict indifference between feature 3 having the value 1 (based on Monica) and feature 3 having the value $-1$ (based on Patricia), i.e., $p$(feature three = 1 | feature one = 1) = $p$(feature 3 = $-1$ | feature one = 1) = 0.5, by Eq. (3). The $multisim_2$ calculations are more complex, for each memory trace, $y$, involving the sum, over all memory traces $z$, of the product of $multisim_1$(probe, $z$) and $multisim_1(z,y)$ [see Eq. (6)]. Here the multiple-step similarity function makes different predictions for Monica and Patricia, because the two-step similarity to Monica, $3s^2 + 2s^4$, is greater than the two-step similarity to Patricia, $s^2 + 3s^3 + s^4$, for all $s$, $0 < s < 1$. By Eq. (3), the subject will predict that feature 3 is more likely to have the value 1 than $-1$. (The degree of this preference depends on $s$.)

This example also illustrates how more complex judgments would be made. If the stimuli were composed of more than three features, and the subject was given two or more feature values as initial information, then the judgments would be modeled the same way, except that the *sim* function would now consider more features.

## EXPERIMENT 1

This first experiment was a simple test of whether people make two-step inferences, if they can infer the value of any stimulus feature, and whether their responses can be accounted for by the multiple-step context model. The domain of the stimuli, social groups, was familiar to the subjects, and subjects made simple judgments between just two features at a time. Subjects studied feature descriptions of 30 fictional people.

After the training on these examples, subjects judged the relations between all possible pairs of feature values, such as the probability that if someone is athletic then he is also married. The test questions allowed subjects to make both one-step and two-step inferences.

## *Method*

*Stimuli.* Table 2 shows the structure of the 30 stimuli, including the frequency of each different description. Each stimulus corresponds to one fictional person who was assigned a male first name. Note that different names could have exactly the same description. The purpose of varying the frequency of descriptions was to have a rich set of stimuli with different conditional relations between different features. In this experiment, for example, feature 1 is a better predictor of feature 2 than it is of feature 3 or 4. The stimuli are described schematically by the values of four binary features. Each feature could take on one of two values, indicated here by 1 and −1. Each person had two feature values known and two feature values missing (indicated by ?'s). When the stimuli were presented to the subjects, the first feature was described as group membership, and could take on the values "Jets" or "Sharks." The three other features were labeled with these pairs of traits: married/single, liberal/conservative, athletic/unathletic. The traits were expected to not evoke strong prior associations from the subjects. Each subject studied stimuli with the same structure, but with different random assignments of names to descriptions, trait pairs to features two through four, and labels to the values 1 and −1 for all four features. These random pairings were intended to further reduce the effects of prior associations on the average responses. Table 3 shows an example of a complete stimulus set shown to one subject, which follows the schema in Table 2.

The training stimuli were constructed so that, in any single example, a value for group membership (Jet or Shark) was paired with just one personal trait (from feature 2, 3, or 4). So, personal traits were directly associated with group membership in the individual stimuli, and it would be possible to infer a trait from a group, or vice versa, by a one-step inference.

TABLE 2
Schematic Description of Training Stimuli for Experiments 1 and 2

| | Features | | | |
|---|---|---|---|---|
| | 1 | 2 | 3 | 4 |
| Frequency | Values | | | |
| 5 | 1 | 1 | ? | ? |
| 5 | −1 | −1 | ? | ? |
| 4 | 1 | ? | 1 | ? |
| 1 | 1 | ? | −1 | ? |
| 4 | −1 | ? | −1 | ? |
| 1 | −1 | ? | 1 | ? |
| 3 | 1 | ? | ? | 1 |
| 2 | 1 | ? | ? | −1 |
| 3 | −1 | ? | ? | −1 |
| 2 | −1 | ? | ? | 1 |
| 30 | | | | |

*Note.* A '?' represents a missing value.

TABLE 3
Sample Training Stimuli for One Subject in Experiment 1

| Name | Description | |
|---|---|---|
| Leonard | Jets | Married |
| Albert | Jets | Married |
| David | Jets | Married |
| Dan | Jets | Married |
| John | Jets | Married |
| Phil | Sharks | Single |
| Lee | Sharks | Single |
| Mark | Sharks | Single |
| Gordon | Sharks | Single |
| Roger | Sharks | Single |
| Herb | Jets | Liberal |
| Amos | Jets | Liberal |
| Misha | Jets | Liberal |
| Brian | Jets | Liberal |
| Jeff | Jets | Conservative |
| Chip | Sharks | Conservative |
| Todd | Sharks | Conservative |
| Steve | Sharks | Conservative |
| Michael | Sharks | Conservative |
| Allen | Sharks | Liberal |
| Scott | Jets | Athletic |
| Ben | Jets | Athletic |
| Robert | Jets | Athletic |
| Tom | Jets | Unathletic |
| Richard | Jets | Unathletic |
| Derek | Sharks | Unathletic |
| Martin | Sharks | Unathletic |
| Kenny | Sharks | Unathletic |
| Frank | Sharks | Athletic |
| Chris | Sharks | Athletic |

To judge the relation between any two traits, such as married and liberal, a subject would have to perform a two-step inference, because individual descriptions contained information about just one trait, with the others unknown. For example, the probe "married" might remind the subject of several members of the Jets, whose political preferences are unknown. Then these Jets, by similarity to other Jets who are liberal, might lead the subject to infer that a married person is also likely liberal.

The test stimuli were 48 ordered pairings of feature values, including all possible pairings except a feature value with itself and a feature value with its opposite. For example, subjects judged the likelihood of "liberal" given all other feature values except "liberal" and "conservative." Each test question presented the task of judging the likelihood of one feature value given an incomplete description consisting of just one other feature value. (The exact wording was "IF someone has the description *X* THEN how likely is he to have the description *Y*?") Twenty-four questions permitted one-step inferences, in which an inference is made from a personal trait to group membership, or vice versa. These questions related feature 1 to feature 2, feature 3, or feature 4. For example, one question asked

subjects to estimate the likelihood of being single given that a person is a Shark. The other 24 questions required a two-step inference to predict one trait from another. These items involved inferences among features 2, 3, and 4, but not feature 1. For example, one question asked how likely someone is to be single if he is unathletic.

*Subjects.* Twenty-five subjects were recruited from two sources: a pool of Stanford University undergraduates taking Introductory Psychology, who received course credit for participation, and undergraduates who responded to an advertisement and received $5.00 for participation. In later experiments, the same recruiting methods were used. No student took part in more than one experiment.

*Procedure.* Subjects were run alone or in groups of two or three. To facilitate memorization, first the subjects examined typed lists of the 30 names and 4 pairs of features. It was explained that for each fictional person, his group was known as was one trait, but no information would be available about the values of the other two traits. It was stressed that each individual did have values on all four features. Each subject performed the experiment on an IBM PC-XT computer running the MEL software system (Schneider, 1988) for stimulus display and data collection. Subjects proceeded at their own pace. In the leaning phase, the subjects first studied the 30 descriptions (e.g., "John is a Jet and liberal"), presented one at a time. Then they were tested on all 60 feature values for the 30 fictional descriptions. Subjects had to choose between a pair of opposite feature values, after being cued by a name (e.g., "Is John liberal or conservative?"). Subjects completed four blocks of this study-and-test cycle. In this as well as all other procedures described for these experiments, all stimulus presentations and test questions were presented in a different random order for each subject.

Next, in the judgment phase, subjects estimated conditional probabilities (on a scale of 0 to 100) in response to the 48 test questions, based on the sample of 30 descriptions that they studied. The subjects were told to ignore any knowledge they had about real people outside the experiment. Subjects were reminded that each fictional person really had three traits, one of each pair of feature values, even though two traits had not been shown. They were encouraged to infer the missing trait information for individual examples. The whole experiment lasted about an hour.

### *Results and Discussion*

The first result is how well the subjects learned the descriptions. Table 4 shows that subjects made steady progress on this memorization task, reaching an average of about 90% on the fourth test cycle (where chance performance is 50%).[3] The results of the judgment phase are of central importance. To analyze the judgment phase responses, the average probability estimate was computed for each of the 48 test questions. In these experiments, the stimuli are labeled schematically such that the same values on different features are associated. For example, Table 2 shows that a value of "1" on feature 1 is predictive of a "1" on feature 2. Likewise, a "−1" on feature 1 is predictive of a "−1" on feature 2. If subjects are sensitive to the relations between features, then they should give responses greater than .5 for "same" judgments: judgments of the probability that some feature has the value $v$ given that some other feature

[3] Dropping the subjects who performed worst on the memorization test did not substantially affect the analyses of the judgment phase, for any experiment.

TABLE 4
Proportion Correct on Training Tests, Experiments 1–5

| Experiment | Block 1 | Block 2 | Block 3 | Block 4 | Range on last test |
|---|---|---|---|---|---|
| 1 | 0.66 | 0.74 | 0.85 | 0.89 | 0.58–1.0 |
| 2 | 0.64 | 0.79 | 0.88 | 0.90 | 0.55–1.0 |
| 3 | 0.74 | 0.87 | 0.93 | | 0.75–1.0 |
| 4 | 0.79 | 0.91 | 0.97 | | 0.77–1.0 |
| 5 | 0.71 | 0.84 | 0.91 | | 0.75–1.0 |

has the same value $v$. Responses greater than .5 would be expected for "same" judgments involving one-step inferences, such as between features 1 and 2, as well as for "same" judgments involving two-step inferences, such as between features 2 and 3. Similarly, if subjects are responding to the conditional relations between features, then they should give responses of less than .5 for "opposite" judgments, such as the probability that feature 2 has the value 1 given that feature 4 has the opposite value, $-1$. It is also expected that for comparable items, such as for all the one-step questions, "same" judgments will be greater than "opposite" judgments.

Table 5 contains representative results, chosen to illustrate performance on different kinds of test questions. For example, the first line shows that when feature 2 had the value 1, the estimated probability of feature 1 having the same value was .72. For the one-step judgments, the "same" responses are well above .5 and the "opposite" responses are below .5. Over all the one-step judgments, the average "same" response was .66 and the average "opposite" response was .37, this difference

TABLE 5
Sample Observed and Predicted Probability Judgments for Experiment 1

| Given | | Judged | | Observed | Model prediction | |
|---|---|---|---|---|---|---|
| Feature | Value | Feature | Value | probability | 2-Step | 1-Step |
| | | | One-step judgments | | | |
| 2 | 1 | 1 | 1 | .72 | .71 | .72 |
| 2 | 1 | 1 | −1 | .29 | .29 | .28 |
| 1 | 1 | 3 | 1 | .69 | .63 | .63 |
| 1 | −1 | 3 | 1 | .35 | .37 | .37 |
| | | | Two-step judgments | | | |
| 2 | 1 | 3 | 1 | .66 | .63 | .50 |
| 2 | 1 | 4 | 1 | .60 | .55 | .50 |

being significant at the $p < .001$ level, paired-$t(24) = 5.91$, $SD = .24$. Thus subjects were able to make one-step judgments.

More interestingly, Table 5 suggests that subjects were also performing two-step inferences, responding greater than .5 to "same" items. The average response for all same-value two-step questions was .58; the average response for opposite-value questions was .47. These two samples were significantly different, paired-$t(24) = 3.35$, $p < .01$, $SD = .17$. Thus, subjects responded as expected on two-step judgments.

Next, both the two-step context model and the original, one-step context model were fitted to the 48 average responses. (Appendix Table A-1 includes the average observed responses, as well as the predictions of the two-step model, for all test questions.) Using the criterion of least squares, values were selected for the four attentional parameters of each model, according to Chandler's (1965) STEPIT algorithm. For features 1 through 4, as defined structurally in Table 2, the best-fitting parameters for the two-step model were .203, .075, .056, and .047, respectively. For the one-step model, the best-fitting parameters were .380, .179, .115, and .043. Next the models' predictions were compared to the average responses. Over the 48 test questions, the correlation between the two-step model's predictions and the observed responses was .976. The mean squared error of this model, defined as the average squared deviation between model and data over all the test questions, was .00119. The performance of the one-step model was clearly worse: $r = .869$ and $MS_e = .00386$. Table 5 also shows representative predictions of each model. Both models give reasonable accounts of one-step inferences. The key to the poor fit of the one-step context model is that it fails to predict any sensitivity to two-step inferences; this model predicts a response of .5 to every two-step judgment, both "same" and "opposite" items. Since subjects in these experiments consistently made two-step inferences, the one-step model will not be considered further in this paper. Figure 1 shows that the two-step model made good predictions overall for both one-step and two-step inferences.[4]

It is clear that in Experiment 1, the extended context model was successful at accounting for people's reasoning from examples. The results showed that people memorized the examples, that people were sensitive to two-step inferences, and that the two-step model described the average responses for both one-step and two-step questions.

[4] Inspection of these figures suggests a slight, but consistent, underprediction by the model. Indeed, over all the probability judgment experiments in this paper, the average response was .52, while the average model prediction was .5. No theoretical interpretation is offered for this response bias on the part of the subjects. However, this bias could easily be "accounted for" by adding a fifth free parameter to the model, with a value of .02.

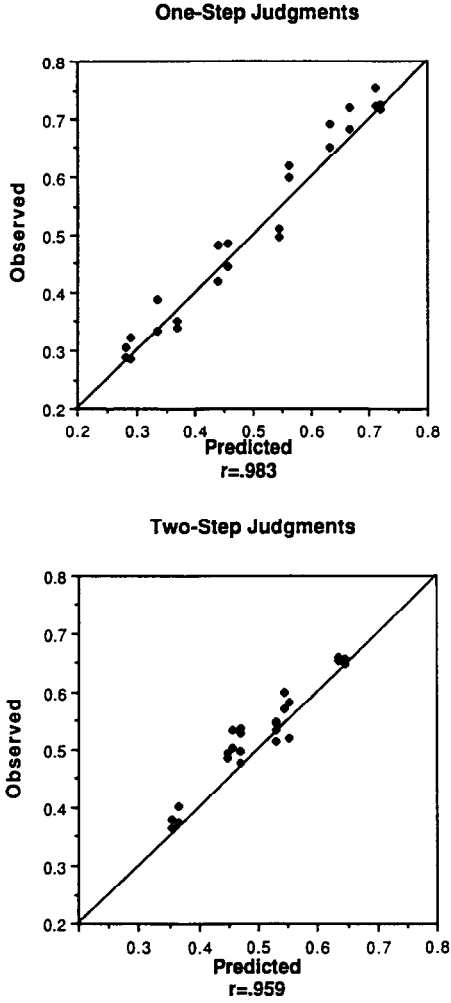


FIG. 1. Observed versus predicted judgments for Experiment 1.

## EXPERIMENT 2

This experiment included more complex test items than Experiment 1, giving two feature values and asking the subject to predict a third feature value. The questions here are whether subjects still make two-step judgments on these more complex questions, and whether the multiple-step context model can describe their performance. Also, to provide some generality, each feature value was identified as a medical symptom, instead of identifying one feature as a group and the others as traits, as in Experiment 1. While the labels in Experiment 1 highlighted the role of the

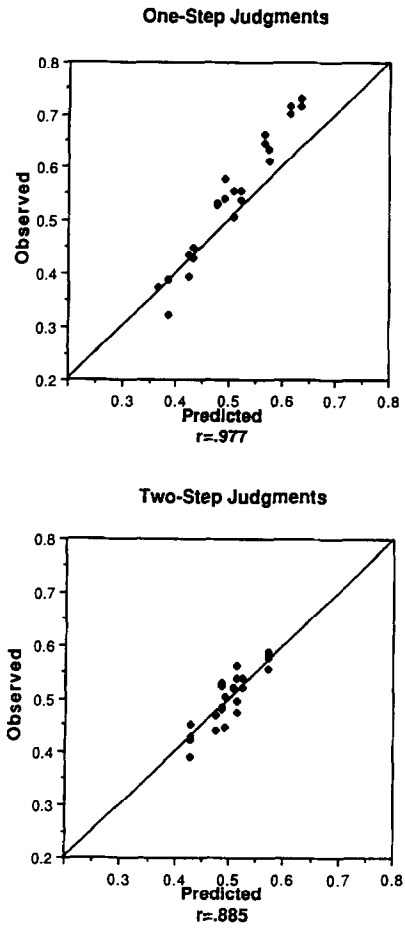


FIG. 2. (A) Observed versus predicted judgments for Experiment 2, one-cue test items. (B) Observed versus predicted judgments for Experiment 2, two-cue test items.

first feature as a category label, the labels in Experiment 2 did not make this distinction. Thus Experiment 2 was intended to demonstrate that subjects can make predictions in the absence of obvious category labels.

## *Method*

*Stimuli.* The 30 training stimuli in Experiment 2 were structured according to Table 2, as in Experiment 1. The 30 descriptions were presented as in Experiment 1, except that all four pairs of feature labels were random permutations of these medical symptoms: high red blood cell count/high white blood cell count, puffy eyes/sunken eyes, weight loss/weight gain, stiff muscles/muscle spasms. As in Experiment 1, each subject got a different assignment of symptoms to the features in Table 2.

The judgment phase consisted of 144 questions. These questions included 48 probability

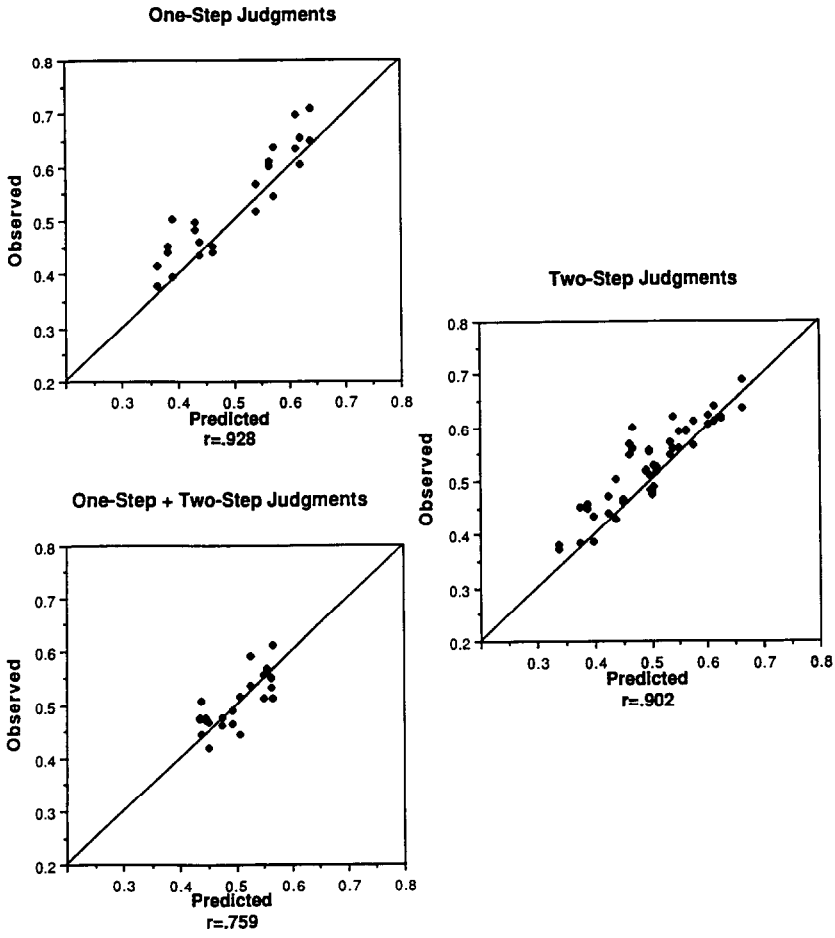


FIG. 2—*Continued*

estimations with the same structure as those in Experiment 1, as well as 96 judgments about the likelihood of one symptom given the values of a *pair* of symptoms. For example, one question asked for the probability of having a high red blood cell count given that someone has puffy eyes and stiff muscles. All possible pairs and triples of symptoms were used except for a symptom grouped with itself and a symptom grouped with its opposite.

*Procedure.* The 36 subjects followed the same procedure used in Experiment 1.

### *Results and Discussion*

Table 4 shows the subjects' average performance on the training blocks. By the fourth block, subjects again averaged about 90% correct. Again using a least-squares criterion to fit the model to the subjects' average responses in the judgment phase, the best estimates of the attentional parameters for the features 1 through 4, as defined in Table 2, were

TABLE 6
Sample Observed and Predicted Probability Judgments for Experiment 2

| Given | | Given | | Judged | | Observed | 2-step model |
|---|---|---|---|---|---|---|---|
| Feature | Value | Feature | Value | Feature | Value | probability | prediction |
| One-step judgments, one cue given | | | | | | | |
| 2 | 1 | | | 1 | 1 | .72 | .63 |
| 3 | 1 | | | 1 | 1 | .61 | .58 |
| Two-step judgments, one cue given | | | | | | | |
| 2 | 1 | | | 3 | 1 | .58 | .57 |
| 4 | −1 | | | 3 | 1 | .45 | .49 |
| One-step judgment, two cues given | | | | | | | |
| 2 | 1 | 3 | 1 | 1 | 1 | .65 | .64 |
| One- plus two-step judgment, two cues given | | | | | | | |
| 1 | 1 | 2 | 1 | 3 | 1 | .61 | .61 |
| Two-step judgments, two cues given | | | | | | | |
| 2 | 1 | 4 | 1 | 3 | 1 | .53 | .56 |
| 2 | 1 | 4 | −1 | 3 | 1 | .57 | .56 |

.354, .141, .146, and .283 respectively. Over the 144 average responses, the correlation between the data and the model's prediction was .893, and the mean squared error was .00195. Figures 2A and 2B show the data plotted against the model's predictions, for judgments when one cue was

TABLE 7
Schematic Description of Training Stimuli for Experiment 3

| | Features | | | |
|---|---|---|---|---|
| | 1 | 2 | 3 | 4 |
| Frequency | Values | | | |
| 4 | 1 | 1 | ? | ? |
| 4 | −1 | −1 | ? | ? |
| 4 | ? | 1 | 1 | ? |
| 4 | ? | −1 | −1 | ? |
| 3 | ? | ? | 1 | 1 |
| 1 | ? | ? | 1 | −1 |
| 3 | ? | ? | −1 | −1 |
| 1 | ? | ? | −1 | 1 |
| 2 | 1 | ? | ? | 1 |
| 2 | 1 | ? | ? | −1 |
| 2 | −1 | ? | ? | −1 |
| 2 | −1 | ? | ? | 1 |
| 32 | | | | |

*Note.* A '?' represents a missing value.

given and for judgments when two cues were given. Overall, the two-step context model performed well for both kinds of judgments. (Appendix Table A-2 includes the average observed responses, as well as the predictions of the two-step model, for all test questions.) Table 6 shows representative average responses, and model predictions, for different kinds of questions. For the one-cue questions, in which subjects made their predictions based on a single feature value, subjects made both one- and two-step judgments as predicted, with "same" responses above .5 and "opposite" responses below .5. Some two-cue test questions involved only one-step judgments, such as predicting the value of feature 1 from features 2 and 3. Other two-cue test questions allowed mixing of two kinds of information, such as predicting the value of feature 3 from feature 1 (a one-step inference) and feature 2 (a two-step inference). Here the model correctly predicts that the response will be higher for this judgment than for the corresponding one-cue, two-step judgment (i.e., predicting feature 3 from just feature 2). Finally, for two-cue, two-step judgments, the model correctly predicts that inferences about feature 3 will be influenced more by feature 2 than by feature 4. As shown in Table 2, feature 4 is the least predictive feature, and changing the value of feature 4 from 1 to −1 does not decrease subjects' judgment of the probability that feature 3 has the value 1. Figures 2A and 2B show that overall the two-step model successfully accounted for these various kinds of judgments.

## EXPERIMENT 3

This experiment was an attempt to generalize the previous findings by using a different stimulus structure. Experiments 1 and 2 highlighted the role of feature 1 as a category label by making its value known in each description, but Experiment 3 did not emphasize feature 1 in this way. Experiment 3 also used different frequencies of descriptions, and different conditional relations between features, than Experiments 1 and 2. Additionally, Experiment 3 provided further generality by using a different response method: forced-choice predictions rather than probability estimates. Collecting binary responses allowed the use of a chi-square goodness-of-fit test to evaluate the model with a hypothesis-testing statistic, in addition to the correlation and mean-squared-error descriptive statistics used so far.

### *Method*

*Stimuli.* Unlike the previous experiments, in Experiment 3, each feature was missing its value for half the fictional person descriptions. Additionally, no description contained the values for both features 1 and 3, or both features 2 and 4, so that two-step inferences would be required for predictions within these pairs. The structure of these stimuli is shown in Table 7. Features 1 and 2 were perfectly predictive of each other, as were features 2 and 3. Features 3 and 4 were strongly correlated. Although values for features 1 and 4 appeared in the same descriptions, these features were uncorrelated. The 32 examples were randomly

assigned female names, and each feature was randomly assigned feature values from these pairs: tall/short, married/single, blonde/brunette, employed/unemployed.

The judgment phase consisted of 72 forced-choice questions. For each question, the subject had to judge which value of a feature is more likely, given an incomplete description consisting of one or two other traits (there were 24 one-cue questions and 48 two-cue questions). For example, subjects were asked "IF a person is tall and single, THEN is she more likely to be employed or unemployed?"

*Procedure.* For the 31 subjects in this experiment, the procedure was identical to Experiment 2 with the following exceptions. First, the learning phase consisted of three study-and-test cycles rather than four, since subjects had not shown much learning during the fourth block in prior experiments. Second, in the judgment phase, subjects answered forced-choice questions about which of a pair of feature values was more likely, rather than estimating the probability of a single feature value.

### *Results and Discussion*

The results of the learning phase are summarized in Table 4. By the third block of tests, the average proportion correct was over 90%.

The two-step context model was fitted to average choice proportions for the 72 judgments, and the best fit had these attentional weights on features 1 through 4: .114, .091, .091, and .153. Once again, the model accounts for the data well. Figure 3 shows the data plotted against the model's predictions. (Heit, 1990 includes the average observed choice proportions, as well as the predictions of the two-step model, for all test questions.) Overall, the correlation between the model and the data is .904, and the mean squared error of the fit is .00747. In addition to the correlation and error statistics, a chi-square goodness of fit test was used

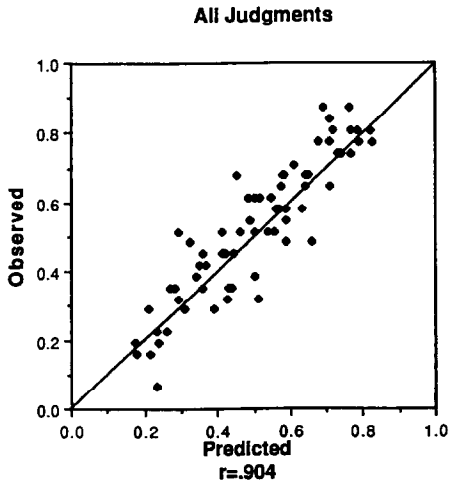


FIG. 3. Observed versus predicted proportions for Experiment 3.

to evaluate the model. Over the 72 test items, the deviations between the data and the four-parameter model were not found to be significant, $\chi^2$ $(67, n = 31) = 75.0, p > .25$, so the multiple-step context model could not be rejected as an account for this data. It is clear that the two-step context model accounts for sensitivity to both one-step and two-step inferences from one cue or two cues given, for both probability judgment and forced-choice responses.

## EXPERIMENT 4

This experiment was intended to answer two questions. First, are people's probability judgments influenced solely by the correct conditional probability, or are they also influenced by the number of examples that the judgments are based on? Relevant to this question is a study by Nisbett, Krantz, Jepson, and Kunda (1983), in which peoples' judgments were sensitive to the sample size. Their subjects were told that 1, 3, or 20 members of a category had a certain property; then the subjects estimated what proportion of all category members had the property. For example, after learning that the 3 people in a sample from a South Pacific tribe were all obese, they estimated that about 55% of the whole tribe would be obese. But after learning that all the members of a 20-person sample were obese, they estimated 75% of the population would be obese. (Nisbett et al. (1983) obtained different results with different categories and properties, but the condition described is most comparable to my stimuli.) Experiment 4 varied sample size for different pairs of features, to see how it affects people's probability judgments, and to see if the multiple-step context model can account for their behavior with different sample sizes.

The second question is whether people can perform inferences of longer than two steps. Experiments 1 through 3 have shown that people make two-step inferences, so it is important to test the boundaries of this phenomenon. Case-based reasoning algorithms do not limit the sequence of retrieval of past examples to two steps, suggesting that people might find it useful to perform longer chains of reasoning. In psychological experiments on memory rather than reasoning, some priming studies have suggested that, in identification and lexical decision tasks, people may be sensitive to multiple-step chains of similarity. For example, "stripes" could prime "tiger," which then primes "lion," making the written word "lion" easier to identify (McNamara & Altarriba, 1988; Ratcliff & McKoon, 1988). The existence of multiple-step priming effects is controversial, but it is uncontroversial that any such effects would be small, because the effect would weaken at each link of the chain. Thus it may be difficult to show three-step chains of inference in an experiment, due to the potential small size of the effect. Nonetheless, Experiment 4 gave

people the opportunity to make three-step inferences, and tested the ability of the $multisim_k$ function with $k = 2$ and $k = 3$ to account for these inferences.

## Method

*Stimuli.* The structure of the 24 training stimuli is shown in Table 8. Note that features 1 and 2 are perfectly predictive of each other, as are features 3 and 4. In the training set, 12 examples contain information about features 1 and 2, while only 4 examples have values for features 3 and 4. This difference in frequency of descriptions instantiates the sample size manipulation. The correct conditional probabilities relating values of features 1 and 2 are all 1.0 or 0.0. For example, the probability of feature 2 having value 1 given that feature 1 has value 1 is 1.0. Likewise, other examples maintain conditional probabilities of 1.0 or 0.0 relating values of feature 3 and feature 4. If subjects are responding only to the correct conditional probabilities, then their estimates would be the same for the two pairs of features. But if people's responses are also sensitive to the sample size, then their estimates would differ, so that they are more conservative when judging the relations between features 3 and 4, compared to judgments about features 1 and 2. Thus, the average probability estimates would be closer to .5 for judgments relating features 3 and 4.

The training stimuli also contained eight examples with information about features 2 and 3; however, no examples had values for both features 1 and 4. Feature 4 could be predicted indirectly from feature 1, by way of a three-step inference: from feature 1 to feature 2, from feature 2 to feature 3, and from feature 3 to feature 4. Similarly, the value of feature 1 could be inferred in three steps from feature 4. The 24 examples were randomly assigned female names, and each feature was randomly assigned feature values from these pairs: tall/short, liberal/conservative, blonde/brunette, employed/unemployed. The judgment phase consisted of 144 probability judgments, following the same structure as those in Experiment 2.

*Procedure.* For the 24 subjects in this experiment, the procedure for the learning phase was identical to Experiment 3. The procedure for the judgment phase was identical to Experiment 2.

TABLE 8
Schematic Description of Training Stimuli for Experiment 4

| | Features | | | |
|---|---|---|---|---|
| | 1 | 2 | 3 | 4 |
| Frequency | Values | | | |
| 6 | 1 | 1 | ? | ? |
| 6 | −1 | −1 | ? | ? |
| 3 | ? | 1 | 1 | ? |
| 1 | ? | 1 | −1 | ? |
| 3 | ? | −1 | −1 | ? |
| 1 | ? | −1 | 1 | ? |
| 2 | ? | ? | 1 | 1 |
| 2 | ? | ? | −1 | −1 |
| 24 | | | | |

*Note.* A '?' represents a missing value.

*Results and Discussion*

As shown in Table 4, subjects memorized these stimuli easily; by the third block, average proportion correct was over 95%. Table 9 shows examples of average judgments in this experiment. There were several important analyses of the judgment phase, including the effect of sample size, testing for sensitivity to three-step judgments, and comparing the fits of the two-step and three-step models to the data. To examine the effect of sample size on conservatism in probability judgments, test questions, involving features 1 and 2 and features 3 and 4, were examined. Eight questions directly asked about the relation between features 1 and 2, including four "same" questions and four "opposite" questions. Similarly, four test questions asked about "same" values for features 3 and 4 and four test questions asked about "opposite" values for these features. As seen in Table 8, the judgments between features 1 and 2 were based on 12 stimuli, while the judgments between features 3 and 4 were based on just 4 stimuli. The 12-stimulus judgments averaged .77 and .26 for "same" and "opposite" items, respectively, while the 4-stimulus judgments averaged .67 and .36 for the "same" and "opposite" items. A two-way analysis of variance (ANOVA) with number of stimuli and same/opposite as factors showed the expected main effect that "same" judgments were greater than "opposite," $F(1,23) = 85.92$, $p < .001$, $MS_e = .049$, and the number of stimuli did not have a significant main effect, $F(1,23) = .16$, $MS_e = .005$. Most important is the significant interaction between the two factors, $F(1,23) = 7.65$, $p < .05$, $MS_e = .030$. Thus subjects' judgments were more sensitive for features 1 and 2 than for features 3 and 4.

Memory differences cannot explain these judgment differences; judg-

TABLE 9
Sample Observed and Predicted Probability Judgments for Experiments 4 and 5

| Given | | Judged | | Experiment 4 | | | Experiment 5 | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | Observed | Model | | Observed | Model | |
| Feature | Value | Feature | Value | Probability | 2-step | 3-step | probability | 2-step | 3-step |
| | | | | One-step judgments | | | | | |
| 4 | 1 | 3 | 1 | .64 | .64 | .63 | .62 | .56 | .56 |
| 3 | 1 | 2 | 1 | .61 | .53 | .54 | .62 | .53 | .52 |
| 2 | 1 | 1 | 1 | .75 | .68 | .66 | .69 | .66 | .64 |
| | | | | Two-step judgments | | | | | |
| 4 | 1 | 2 | 1 | .51 | .53 | .57 | .48 | .52 | .54 |
| 3 | 1 | 1 | 1 | .59 | .55 | .57 | .57 | .54 | .55 |
| | | | | Three-step judgment | | | | | |
| 4 | 1 | 1 | 1 | .51 | .50 | .52 | .57 | .50 | .51 |

ments on features 3 and 4 were not less sensitive due to worse memory for these features. To test whether some features were learned better than others, a two-way ANOVA was performed, with feature number (as defined in Table 8) and training block number as factors. Feature number was found to be a significant factor, $F(3,69) = 3.81, p < .05, MS_e = .016$, also, block number was a significant factor, $F(2,46) = 51.15, p < .001$, $MS_e = .015$. The mean proportions correct on features 1 through 4 were .86, .90, .88, and .93. Using Tukey's HSD test, the only conclusion to be drawn is that memory for feature 4 is better than memory for feature 1, at the $p < .05$ level.

Next the responses were examined for evidence of three-step inferences. Eight one-cue test questions involved just features 1 and 4. Of these three-step items, the four "same" questions had an average response of .51, and the four "opposite" questions had an average response of .49. While this difference is in the expected direction, a paired-$t$ test showed the difference to not be significant, $t(23) = .37$, $SD = .22$. This analysis of one-cue test questions did not detect use of three-step inferences. However, the following analyses may be more sensitive in examining whether three-step inferences had an overall effect on all the test questions, both one-cue and two-cue.

The extended context model was applied to the data using two simulations, with the $multisim_k$ function having a subscript of 2 or 3, corresponding to the inclusion of two-step inferences and three-step inferences. It was expected that if people made three-step inferences, then the model using a subscript of 3 would predict the data better. Note that the three-step version of the model makes predictions that also include information about one-step and two-step inferences, just as the two-step model also allows one-step predictions. Table 9 shows predictions of each model on representative test questions. The two-step model predicts no sensitivity, i.e., a response of .5, on all three-step questions. With a $k$ subscript of 2, the best-fitting attentional parameters for the four features were .159, .245, .286, and .102. Over the 144 average responses, the correlation between the data and the model's prediction was .939, and the mean squared error was .00187. With a $k$ subscript of 3, the best-fitting attentional parameters for the four features were .095, .182, .204, and .045, and the fit between the data and the three-step model's prediction was worse, $r = .869$, $MS_e = .00368$. Figure 4 illustrates how the two-step model characterizes the data better overall than the three-step model. (Appendix Table A-3 includes the average observed responses, as well as the predictions of the two-step model, for all test questions. Heit, 1990 also includes predictions of the three-step model.)

To return to sample size effects and the multiple-step context model, it was found that the model did correctly predict the direction of sample size

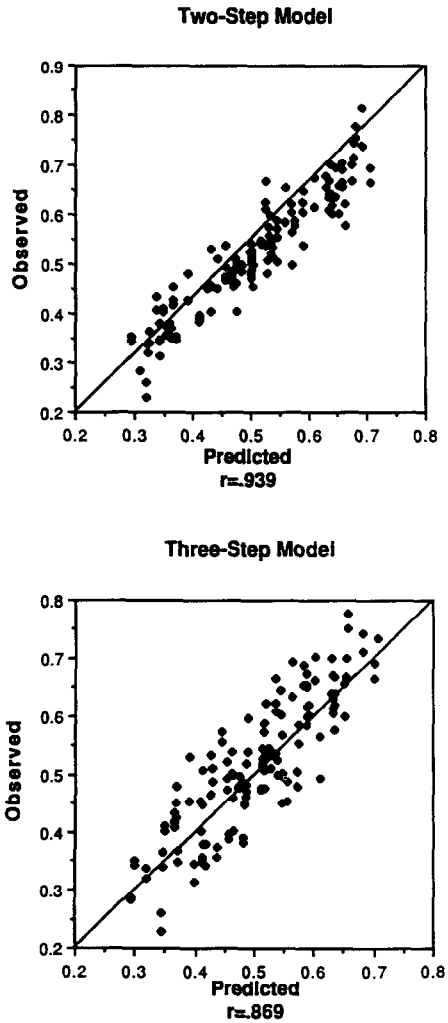


FIG. 4. Observed versus predicted judgments for Experiment 4.

effects. The two-step model's predictions for the 12-stimulus judgments were .68 and .32 for "same" and "opposite" judgments, respectively, while the predictions for the 4-stimulus judgments were .63 and .37. The model, like the subjects, showed less sensitivity to relations between features 3 and 4 than to features 1 and 2. This effect was not caused by higher attentional parameters (i.e., less attention) on features 3 and 4. Instead, the weights for features 1 and 2 (.159 and .245) fell between the weights for features 3 and 4 (.286 and .102). The multiple-step model succeeded because its memory traces interacted with each other during

the second step of inference. The 12-stimulus predictions were less conservative than the 4-step predictions because the 12-stimulus predictions were based on a relatively large set of instances with a high degree of mutual similarity. The 12 memory traces containing values for features 1 and 2 reinforced each other, during the second step of reasoning, more than did the 4 memory traces containing values for features 3 and 4, causing the multiple-step model to predict stronger judgments between features 1 and 2.

To summarize, this experiment demonstrated a sample size bias consistent with Nisbett et al. (1983). The multiple-step context model accounted for this result, in terms of more mutual facilitation within larger sets of memory traces. Although the stimuli and test questions permitted three-step inferences, Experiment 4 did not show any evidence for these inferences. The two-step context model gave a good account of the average responses on all the test questions.

## EXPERIMENT 5

This experiment was a final-comparison between the two-step and three-step models, including questions permitting three-step judgments. While Experiment 4 showed no evidence for three-step chains of inference, we cannot accept that null result at face value. Due to the success of the sample-size manipulation, subjects showed little sensitivity to the relation between features 3 and 4. Thus the apparent lack of three-step inferences might actually be due to these features being a "weak link" in the chain. The training stimuli in Experiment 5 did not include small samples of feature pairings, in an attempt to encourage stronger, less conservative judgments than in Experiment 4, so that three-step inferences might be more evident.

### *Method*

*Stimuli.* The 30 training stimuli are shown schematically in Table 10. These stimuli are similar to those in Experiment 4, except that there are 10 stimuli with values for features 1 and 2, 10 stimuli with values for features 2 and 3, and 10 stimuli with values for features 3 and 4. It would be possible to predict feature 1 from feature 4, or vice versa, by a three-step inference. Features 1 and 2 are perfectly correlated, while features 2 and 3 as well as features 3 and 4 are strongly related. The stimuli were designed with strong relations between adjacent features to make it easier for subjects to perform remote inferences between features 1 and 4. The training stimuli in this experiment were presented as in Experiment 4, and the 144 judgment stimuli had the same structure as those of Experiment 4.

*Procedure.* The 28 subjects in this experiment followed the same procedure as that of Experiment 4.

### *Results and Discussion*

The results of the learning phase are summarized in Table 4. By the third block of tests, the average proportion correct was over 90%. Table

TABLE 10
Schematic Description of Training Stimuli for Experiment 5

| | Features | | | |
|---|---|---|---|---|
| | 1 | 2 | 3 | 4 |
| Frequency | Values | | | |
| 5 | 1 | 1 | ? | ? |
| 5 | −1 | −1 | ? | ? |
| 4 | ? | 1 | 1 | ? |
| 1 | ? | 1 | −1 | ? |
| 4 | ? | −1 | −1 | ? |
| 1 | ? | −1 | 1 | ? |
| 4 | ? | ? | 1 | 1 |
| 1 | ? | ? | 1 | −1 |
| 4 | ? | ? | −1 | −1 |
| 1 | ? | ? | −1 | 1 |
| 30 | | | | |

*Note.* A '?' represents a missing value.

9 shows data from representative test questions. As in Experiment 4, one-cue questions from the judgment phase were examined for evidence of three-step inferences between features 1 and 4. The average judgment for "same" questions was .55, and the average for "opposite" questions was .50. Again, the direction of this difference is suggestive of three-step inferences, but a paired-$t$ test on "same" versus "opposite" judgments showed that the difference was not significant, $t(27) = .95$, $SD = .27$. Thus on the 8 single-cue questions permitting three-step inferences, subjects did not show a reliable trend toward making these inferences.

The role of three-step inferences was assessed over all 144 questions, by fitting the two-step model and the three-step model to the data. For the two-step model, the best-fitting attentional parameters for the four features were .153, .266, .353, and .203. The correlation between the data and the model's predictions was .918, $MS_e = .00160$. For the three-step model, the best-fitting attentional parameters were .097, .188, .320, and .103; the correlation between data and the three-step model's prediction was .864, $MS_e = .00242$. Overall, the two-step model performs better. Figure 5 compares how the two-step model and the three-step model fit the data on one- and two-cue test questions involving three-step inferences (those questions with feature 1 as a cue and feature 4 to be predicted, or vice versa) and how they fit the data on the remaining questions permitting only one- and two-step inferences (those questions which do not have feature 1 or 4 as a cue, or do not involve the prediction of feature 1 or 4). The three-step model does account for the three-step inferences

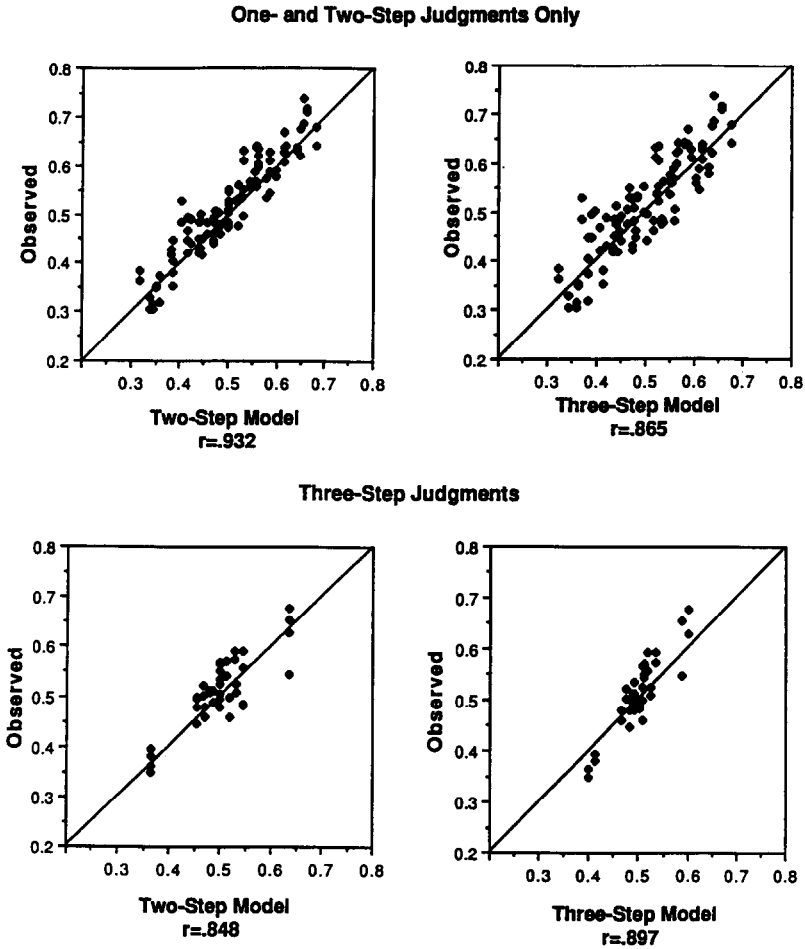


FIG. 5. Observed versus predicted judgments for Experiment 5.

somewhat better than the two-step model. (Appendix Table A-3 includes the average observed responses, as well as the predictions of the two-step model, for all test questions. Heit, 1990 also includes predictions of the three-step model.)

In sum, most subjects on most test questions were not influenced by three-step inferences. Indeed, on the one-cue three-step judgments, there was no significant difference on judgments of "same" versus "opposite" values. However, over all the three-step questions, the three-step model does appear to describe the data slightly better than the two-step model. Thus some subjects, on some test questions, may have been influenced by three-step inferences. To obtain convincing evidence for three-step judg-

ments might require using stimuli that lead people to expect perfect, rather than probabilistic, dependencies between features. With probabilistic feature relations, any differences between responses to various three-step questions may be too small to detect. In fact, Griggs (1976) ran several experiments with perfectly deterministic stimuli, that still showed weak effects of multiple-step inferences. He presented subjects with sentences of the form "All As are Bs; All Bs are Cs; All Cs are Ds." Over various experimental manipulations, Griggs found that only between 55 and 80% of subjects accepted the three-step inference that All As are Ds.

## GENERAL DISCUSSION

### *Summary*

These five experiments have shown that people can use memory for examples to perform complex reasoning tasks. These reasoning tasks include accurately predicting the value of any feature of a stimulus, using sample size to moderate these predictions, and making two-step inferences from chains of similar examples. People may also be able to perform inferences with chains longer than two steps, when the relations between features are nearly deterministic rather than probabilistic. Experiments involving all of these reasoning tasks were accounted for using a mathematical model incorporating memory for examples and a simple similarity function with a chaining operation. I will now offer some critical evaluation of the experiments and theory presented here, and compare this work to that of some other researchers.

### *Limitations of Approach*

The approach taken in this paper has both advantages and disadvantages. By closely controlling the structure of the examples that people memorized, it was possible to make precise predictions about the subjects' judgments. By selecting stimulus labels to avoid content effects and prior associations, it was possible to assess the role that similarity based on feature overlap has on reasoning; and to do so without considering the role of theories or other background knowledge. These advantages are common to experiments in the artificial category learning paradigm (e.g., Brooks, 1978; Medin & Schaffer, 1978; Nosofsky, 1984).

These advantages come with matching disadvantages. First, the artificial category paradigm uses simple stimuli that can make it difficult for researchers to draw conclusions about natural categories and real-world reasoning. For example, the experiments here do not capture the idea that some features are labels for natural kinds, or that some properties for some categories are inherently less predictable than others.

The second disadvantage is that by isolating similarity, we may fail to

see the bigger picture of reasoning, when more factors, such as theories and background knowledge are involved in the process. The multiple-step context model assumes that instances are stored, retrieved, and compared using a simple similarity function without reference to other knowledge. Murphy and Medin (1985) and others have argued that similarity is not sufficient to explain the structure of people's categories; rather background knowledge and causal theories about categories must be considered as well. The present experiments may best reflect similarity-based reasoning that is subsequent to any effects of theories.

*The Multiple-Step Context Model*

It is clear that the two-step context model was able to account for the results of these five experiments. The breakdowns of test items showed that the model predicted people's one-step and two-step inferences, while the original context model (Medin & Schaffer, 1978) was only able to account for the one-step inferences. Additionally, the two-step model was successful at predicting how subjects used the values of two different features to make predictions about a third. This process of combining the evidence from two or more (perhaps conflicting) cues to make a probability judgment was not highlighted in the analyses presented here. Yet the model makes definite predictions that could be tested with sufficiently sensitive experiments. For example, in Experiment 5, the value of feature 1 could be predicted by a one-step inference from the value of feature 2, and by a two-step inference from the value of feature 3. If the values of both features 2 and 3 are given, the two-step model predicts that both of these pieces of information affect judgments about feature 1. An alternate model could predict that only feature 2 would affect judgments about feature 1, because subjects would end the reasoning process after a successful one-step inference. Subjects might only perform two-step inferences when one-step inferences are not informative.

Also, the multiple-step context model has been presented with a description of sequential memory retrieval used to make inferences at the time of testing. However, this story is not inherent to the equations in the model. Instead, the similarity-based chains of reasoning might be formed at the time of encoding of examples. The present experiments support the equations, but do not distinguish between the retrieval and encoding stories.

*Other Models for this Task*

Can other mathematical models explain these results? An abstraction model for these tasks might describe subjects as directly estimating conditional probabilities, rather than reasoning from instance memory traces. To make complex judgments, involving multiple features or between fea-

tures that do not occur directly with each other in the stimuli, the subjects would combine these estimated probabilities according to various statistical rules.

One such model based on estimated probabilities, proposed by Anderson (1991), has accounted for many previous experimental findings in categorization. His "rational model" for category judgments uses estimated probabilities according to Bayesian statistics. The rational model separates the stimuli in an experiment into different partitions by a clustering algorithm that is sensitive to the similarity structure of the stimuli. When this model is presented with a new stimulus with the value of some feature unknown, the model can predict the likelihood that this stimulus has a particular value for the feature. To make this judgment, the rational model uses a sum of products of estimated probabilities. The model sums up, over all the partitions in memory, the estimated probability that the new stimulus belongs in the partition multiplied by the estimated probability that an item in that partition has the feature value of interest. Anderson has not applied his model to tasks involving predicting any feature of a stimulus rather than just the category label, nor has he applied the model to multiple-step inferences.

Using a computer simulation of the rational model, supplied by Anderson, I attempted to account for the results in this paper. In its simplest form, the rational model has just a single parameter, $c$, that affects how many partitions are formed. After informal sampling of the parameter space, I selected a value, $c = 0.4$, that allowed the rational model to predict the results of the experiments quite well. Because the rational model uses a probabilistic partitioning algorithm, I ran 250 simulations for each experiment and used the average predictions for analysis. The rational model only makes ordinal predictions about probabilities, but for all five experiments, the rank-order correlations between the rational model and the data were comparable to the rank-order correlations between the multiple-step context model and the data (see Table 11).

TABLE 11
Comparison of Rational Model and Two-Step Model for Experiments 1–5

| Experiment | Rank-order correlations | |
|---|---|---|
| | Rational model[a] | Two-step model |
| 1 | 0.961 | 0.961 |
| 2 | 0.881 | 0.892 |
| 3 | 0.874 | 0.899 |
| 4 | 0.851 | 0.941 |
| 5 | 0.886 | 0.898 |

[a] Rational model was simulated by 250 runs with coupling parameter $c = 0.4$.

Although not specifically designed for this purpose, the rational model was able to account for two-step inferences. Further simulations showed that the model could also make three-step inferences. The rational model makes these multiple-step inferences without explicitly using a multiple-step reasoning process. Instead, the rational model forms its chains of reasoning in the partitions. In Experiment 1, for example, the model tended to put Jet instances into one partition and Shark instances into another partition, even joining together instances that matched only in terms of group membership. By putting an athletic Jet in the same partition as a married Jet, the clustering algorithm established a two-step link between athletic and married, by way of the overlapping feature value "Jet."

Does the success of the rational model mean that the subjects in these experiments were reasoning by combining conditional probabilities, rather than reasoning directly with instance memory traces? Not necessarily. Anderson described his model not as a plausible process description of what people do, but as a rational, statistical justification for both human judgments and the models that psychologists construct. However, it may be possible to come up with a more plausible process model that also involves probability estimation rather than reasoning directly from instances, and that can account for the results of these experiments. Barsalou (1990) argued that the success of some instance-based model of categorization does not rule out the entire class of possible abstraction models. However, these experiments do rule out certain models; without processing assumptions such as those presented here, the context model (Medin & Schaffer, 1978) as well as other instance models cannot perform the tasks described in this paper.[5] The multiple-step context model makes the contribution of extending our conception of instance-based memory to allow for more complex reasoning.

## CONCLUSION

This research serves as a mathematical treatment and experimental test of some of the philosopher John Stuart Mill's (1874) ideas about reasoning. Mill was particularly interested in reasoning from examples; he asserted that "All inference is from particulars to particulars" (1874, Book

[5] Another computer model for inferring missing features of stimuli from memory for examples was proposed by McClelland (1981). This model uses instance representations like the multiple-step context model, but uses a spreading activation function to make its predictions. This model can predict any stimulus feature and perform chains of reasoning based on similarity; it is likely that McClelland's model could predict results qualitatively consistent with the results of these experiments. However, the McClelland model was presented as a demonstration of connectionist techniques, rather than as a model for fitting psychological data, so it is not tested further here.

II, Chapt. I, Sec. 3). Mill also proposed the fundamental rule of similarity-based reasoning from examples: if an object *A* possesses a certain property *P*, then as we discover greater degrees of similarity between *A* and some object *B*, we are more likely to infer that *B* also has property *P* [1874, Book III, Chapt. XX, Sec. 2; also, cf. Eqs. (2 and 3)].

The two most novel aspects of this paper were also present in Mill's work. First, Mill pointed out, that for the purpose of prediction, any property of an object can be treated as a category label (1874, Book IV, Chapt. VII, Sec. 2). And second, he argued for the prevalence and the usefulness of *trains of reasoning*, which are like two-step and three-step inferences discussed here (1874, Book II, Chapt. IV, Sec. 2).

Mill's work, as well as this investigation, support the conclusion that our memories for the particular events, objects, and people of our experience are not merely passive memory traces used for a limited range of tasks. Instead, our memories for examples form the basis of varied and powerful kinds of reasoning.

## APPENDIX

### Table A-1

*Observed and Predicted Probability Judgments for Experiment 1*

| Given | | Judged | | Observed probability | 2-step model |
|---|---|---|---|---|---|
| Feature | Value | Feature | Value | | |
| 2 | 1 | 1 | 1 | 0.72 | 0.71 |
| 2 | −1 | 1 | 1 | 0.32 | 0.29 |
| 3 | 1 | 1 | 1 | 0.72 | 0.67 |
| 3 | −1 | 1 | 1 | 0.34 | 0.34 |
| 4 | 1 | 1 | 1 | 0.62 | 0.56 |
| 4 | −1 | 1 | 1 | 0.42 | 0.44 |
| 1 | 1 | 2 | 1 | 0.73 | 0.72 |
| 1 | 1 | 2 | −1 | 0.29 | 0.28 |
| 1 | 1 | 3 | 1 | 0.69 | 0.63 |
| 1 | 1 | 3 | −1 | 0.35 | 0.37 |
| 1 | 1 | 4 | 1 | 0.51 | 0.54 |
| 1 | 1 | 4 | −1 | 0.45 | 0.46 |
| 2 | 1 | 1 | −1 | 0.29 | 0.29 |
| 2 | −1 | 1 | −1 | 0.75 | 0.71 |
| 3 | 1 | 1 | −1 | 0.39 | 0.34 |
| 3 | −1 | 1 | −1 | 0.68 | 0.67 |
| 4 | 1 | 1 | −1 | 0.48 | 0.44 |
| 4 | −1 | 1 | −1 | 0.60 | 0.56 |
| 1 | −1 | 2 | 1 | 0.31 | 0.28 |
| 1 | −1 | 2 | −1 | 0.72 | 0.72 |
| 1 | −1 | 3 | 1 | 0.34 | 0.37 |
| 1 | −1 | 3 | −1 | 0.65 | 0.63 |
| 1 | −1 | 4 | 1 | 0.49 | 0.46 |
| 1 | −1 | 4 | −1 | 0.50 | 0.54 |
| 3 | 1 | 2 | 1 | 0.65 | 0.65 |
| 3 | −1 | 2 | 1 | 0.37 | 0.35 |
| 4 | 1 | 2 | 1 | 0.58 | 0.55 |
| 4 | −1 | 2 | 1 | 0.49 | 0.45 |

## Table A-1 *Continued*

| Given | | Judged | | Observed probability | 2-step model |
|---|---|---|---|---|---|
| Feature | Value | Feature | Value | | |
| 3 | 1 | 2 | −1 | 0.38 | 0.35 |
| 3 | −1 | 2 | −1 | 0.66 | 0.65 |
| 4 | 1 | 2 | −1 | 0.49 | 0.45 |
| 4 | −1 | 2 | −1 | 0.52 | 0.55 |
| 2 | 1 | 3 | 1 | 0.66 | 0.63 |
| 2 | −1 | 3 | 1 | 0.37 | 0.37 |
| 4 | 1 | 3 | 1 | 0.55 | 0.53 |
| 4 | −1 | 3 | 1 | 0.54 | 0.47 |
| 2 | 1 | 3 | −1 | 0.40 | 0.37 |
| 2 | −1 | 3 | −1 | 0.65 | 0.63 |
| 4 | 1 | 3 | −1 | 0.50 | 0.47 |
| 4 | −1 | 3 | −1 | 0.55 | 0.53 |
| 2 | 1 | 4 | 1 | 0.60 | 0.55 |
| 2 | −1 | 4 | 1 | 0.53 | 0.46 |
| 3 | 1 | 4 | 1 | 0.51 | 0.53 |
| 3 | −1 | 4 | 1 | 0.48 | 0.47 |
| 2 | 1 | 4 | −1 | 0.50 | 0.46 |
| 2 | −1 | 4 | −1 | 0.57 | 0.55 |
| 3 | 1 | 4 | −1 | 0.53 | 0.47 |
| 3 | 1 | 4 | −1 | 0.54 | 0.53 |

## Table A-2

*Observed and Predicted Probability Judgments for Experiment 2*

| Given | | Given | | Judged | | Observed probability | 2-step model |
|---|---|---|---|---|---|---|---|
| Feature | Value | Feature | Value | Feature | Value | | |
| 2 | 1 | | | 1 | 1 | 0.72 | 0.63 |
| 2 | −1 | | | 1 | 1 | 0.38 | 0.37 |
| 3 | 1 | | | 1 | 1 | 0.61 | 0.58 |
| 3 | −1 | | | 1 | 1 | 0.39 | 0.42 |
| 4 | 1 | | | 1 | 1 | 0.51 | 0.51 |
| 4 | −1 | | | 1 | 1 | 0.58 | 0.49 |
| 2 | 1 | 3 | 1 | 1 | 1 | 0.65 | 0.64 |
| 2 | 1 | 3 | −1 | 1 | 1 | 0.52 | 0.54 |
| 2 | 1 | 4 | 1 | 1 | 1 | 0.60 | 0.62 |
| 2 | 1 | 4 | −1 | 1 | 1 | 0.64 | 0.61 |
| 2 | −1 | 3 | 1 | 1 | 1 | 0.44 | 0.46 |
| 2 | −1 | 3 | −1 | 1 | 1 | 0.42 | 0.36 |
| 2 | −1 | 4 | 1 | 1 | 1 | 0.39 | 0.39 |
| 2 | −1 | 4 | −1 | 1 | 1 | 0.45 | 0.38 |
| 3 | 1 | 4 | 1 | 1 | 1 | 0.55 | 0.57 |
| 3 | 1 | 4 | −1 | 1 | 1 | 0.60 | 0.56 |
| 3 | −1 | 4 | 1 | 1 | 1 | 0.46 | 0.44 |
| 3 | −1 | 4 | −1 | 1 | 1 | 0.50 | 0.43 |
| 2 | 1 | | | 1 | −1 | 0.38 | 0.37 |
| 2 | −1 | | | 1 | −1 | 0.73 | 0.63 |
| 3 | 1 | | | 1 | −1 | 0.43 | 0.42 |
| 3 | −1 | | | 1 | −1 | 0.64 | 0.58 |
| 4 | 1 | | | 1 | −1 | 0.54 | 0.49 |
| 4 | −1 | | | 1 | −1 | 0.56 | 0.51 |
| 2 | 1 | 3 | 1 | 1 | −1 | 0.38 | 0.36 |
| 2 | 1 | 3 | −1 | 1 | −1 | 0.45 | 0.46 |
| 2 | 1 | 4 | 1 | 1 | −1 | 0.44 | 0.38 |

## Table A-2 *Continued*

| Given | | Given | | Judged | | Observed probability | 2-step model |
|---|---|---|---|---|---|---|---|
| Feature | Value | Feature | Value | Feature | Value | | |
| 2 | 1 | 4 | −1 | 1 | −1 | 0.50 | 0.39 |
| 2 | −1 | 3 | 1 | 1 | −1 | 0.57 | 0.54 |
| 2 | −1 | 3 | −1 | 1 | −1 | 0.71 | 0.64 |
| 2 | −1 | 4 | 1 | 1 | −1 | 0.70 | 0.61 |
| 2 | −1 | 4 | −1 | 1 | −1 | 0.66 | 0.62 |
| 3 | 1 | 4 | 1 | 1 | −1 | 0.48 | 0.43 |
| 3 | 1 | 4 | −1 | 1 | −1 | 0.44 | 0.44 |
| 3 | −1 | 4 | 1 | 1 | −1 | 0.61 | 0.56 |
| 3 | −1 | 4 | −1 | 1 | −1 | 0.64 | 0.57 |
| 1 | 1 | | | 2 | 1 | 0.70 | 0.61 |
| 1 | −1 | | | 2 | 1 | 0.39 | 0.39 |
| 3 | 1 | | | 2 | 1 | 0.58 | 0.57 |
| 3 | −1 | | | 2 | 1 | 0.45 | 0.43 |
| 4 | 1 | | | 2 | 1 | 0.50 | 0.51 |
| 4 | −1 | | | 2 | 1 | 0.48 | 0.49 |
| 1 | 1 | 3 | 1 | 2 | 1 | 0.69 | 0.66 |
| 1 | 1 | 3 | −1 | 2 | 1 | 0.56 | 0.55 |
| 1 | 1 | 4 | 1 | 2 | 1 | 0.62 | 0.62 |
| 1 | 1 | 4 | −1 | 2 | 1 | 0.60 | 0.60 |
| 1 | −1 | 3 | 1 | 2 | 1 | 0.46 | 0.45 |
| 1 | −1 | 3 | −1 | 2 | 1 | 0.37 | 0.34 |
| 1 | −1 | 4 | 1 | 2 | 1 | 0.44 | 0.40 |
| 1 | −1 | 4 | −1 | 2 | 1 | 0.45 | 0.38 |
| 3 | 1 | 4 | 1 | 2 | 1 | 0.61 | 0.57 |
| 3 | 1 | 4 | −1 | 2 | 1 | 0.51 | 0.55 |
| 3 | −1 | 4 | 1 | 2 | 1 | 0.47 | 0.45 |
| 3 | −1 | 4 | −1 | 2 | 1 | 0.48 | 0.43 |
| 1 | 1 | | | 2 | −1 | 0.32 | 0.39 |
| 1 | −1 | | | 2 | −1 | 0.72 | 0.61 |
| 3 | 1 | | | 2 | −1 | 0.42 | 0.43 |
| 3 | −1 | | | 2 | −1 | 0.56 | 0.57 |
| 4 | 1 | | | 2 | −1 | 0.48 | 0.49 |
| 4 | −1 | | | 2 | −1 | 0.47 | 0.51 |
| 1 | 1 | 3 | 1 | 2 | −1 | 0.38 | 0.34 |
| 1 | 1 | 3 | −1 | 2 | −1 | 0.47 | 0.45 |
| 1 | 1 | 4 | 1 | 2 | −1 | 0.39 | 0.38 |
| 1 | 1 | 4 | −1 | 2 | −1 | 0.39 | 0.40 |
| 1 | −1 | 3 | 1 | 2 | −1 | 0.59 | 0.55 |
| 1 | −1 | 3 | −1 | 2 | −1 | 0.64 | 0.66 |
| 1 | −1 | 4 | 1 | 2 | −1 | 0.62 | 0.60 |
| 1 | −1 | 4 | −1 | 2 | −1 | 0.62 | 0.62 |
| 3 | 1 | 4 | 1 | 2 | −1 | 0.48 | 0.43 |
| 3 | 1 | 4 | −1 | 2 | −1 | 0.42 | 0.45 |
| 3 | −1 | 4 | 1 | 2 | −1 | 0.56 | 0.55 |
| 3 | −1 | 4 | −1 | 2 | −1 | 0.51 | 0.57 |
| 1 | 1 | | | 3 | 1 | 0.66 | 0.57 |
| 1 | −1 | | | 3 | 1 | 0.45 | 0.43 |
| 2 | 1 | | | 3 | 1 | 0.58 | 0.57 |
| 2 | −1 | | | 3 | 1 | 0.39 | 0.43 |
| 4 | 1 | | | 3 | 1 | 0.52 | 0.51 |
| 4 | 1 | | | 3 | 1 | 0.45 | 0.49 |
| 1 | 1 | 2 | 1 | 3 | 1 | 0.61 | 0.61 |
| 1 | 1 | 2 | −1 | 3 | 1 | 0.56 | 0.50 |
| 1 | 1 | 4 | 1 | 3 | 1 | 0.57 | 0.58 |
| 1 | 1 | 4 | −1 | 3 | 1 | 0.59 | 0.56 |
| 1 | −1 | 2 | 1 | 3 | 1 | 0.53 | 0.51 |
| 1 | −1 | 2 | −1 | 3 | 1 | 0.45 | 0.39 |
| 1 | −1 | 4 | 1 | 3 | 1 | 0.50 | 0.44 |

## Table A-2 *Continued*

| Given | | Given | | Judged | | Observed probability | 2-step model |
|---|---|---|---|---|---|---|---|
| Feature | Value | Feature | Value | Feature | Value | | |
| 1 | −1 | 4 | −1 | 3 | 1 | 0.47 | 0.43 |
| 2 | 1 | 4 | 1 | 3 | 1 | 0.53 | 0.56 |
| 2 | 1 | 4 | −1 | 3 | 1 | 0.57 | 0.56 |
| 2 | −1 | 4 | 1 | 3 | 1 | 0.47 | 0.45 |
| 2 | −1 | 4 | −1 | 3 | 1 | 0.45 | 0.44 |
| 1 | 1 | | | 3 | −1 | 0.43 | 0.43 |
| 1 | −1 | | | 3 | −1 | 0.65 | 0.57 |
| 2 | 1 | | | 3 | −1 | 0.43 | 0.43 |
| 2 | −1 | | | 3 | −1 | 0.59 | 0.57 |
| 4 | 1 | | | 3 | −1 | 0.51 | 0.49 |
| 4 | 1 | | | 3 | −1 | 0.52 | 0.51 |
| 1 | 1 | 2 | 1 | 3 | −1 | 0.46 | 0.39 |
| 1 | 1 | 2 | −1 | 3 | −1 | 0.49 | 0.51 |
| 1 | 1 | 4 | 1 | 3 | −1 | 0.44 | 0.43 |
| 1 | 1 | 4 | −1 | 3 | −1 | 0.43 | 0.44 |
| 1 | −1 | 2 | 1 | 3 | −1 | 0.55 | 0.50 |
| 1 | −1 | 2 | −1 | 3 | −1 | 0.64 | 0.61 |
| 1 | −1 | 4 | 1 | 3 | −1 | 0.59 | 0.56 |
| 1 | −1 | 4 | −1 | 3 | −1 | 0.61 | 0.58 |
| 2 | 1 | 4 | 1 | 3 | −1 | 0.51 | 0.44 |
| 2 | 1 | 4 | −1 | 3 | −1 | 0.48 | 0.45 |
| 2 | −1 | 4 | 1 | 3 | −1 | 0.56 | 0.56 |
| 2 | −1 | 4 | −1 | 3 | −1 | 0.55 | 0.56 |
| 1 | 1 | | | 4 | 1 | 0.56 | 0.52 |
| 1 | −1 | | | 4 | 1 | 0.53 | 0.48 |
| 2 | 1 | | | 4 | 1 | 0.54 | 0.52 |
| 2 | −1 | | | 4 | 1 | 0.44 | 0.48 |
| 3 | 1 | | | 4 | 1 | 0.56 | 0.51 |
| 3 | −1 | | | 4 | 1 | 0.53 | 0.49 |
| 1 | 1 | 2 | 1 | 4 | 1 | 0.62 | 0.54 |
| 1 | 1 | 2 | −1 | 4 | 1 | 0.51 | 0.50 |
| 1 | 1 | 3 | 1 | 4 | 1 | 0.57 | 0.53 |
| 1 | 1 | 3 | −1 | 4 | 1 | 0.52 | 0.51 |
| 1 | −1 | 2 | 1 | 4 | 1 | 0.48 | 0.50 |
| 1 | −1 | 2 | −1 | 4 | 1 | 0.57 | 0.46 |
| 1 | −1 | 3 | 1 | 4 | 1 | 0.52 | 0.49 |
| 1 | −1 | 3 | −1 | 4 | 1 | 0.56 | 0.47 |
| 2 | 1 | 3 | 1 | 4 | 1 | 0.59 | 0.53 |
| 2 | 1 | 3 | −1 | 4 | 1 | 0.52 | 0.51 |
| 2 | −1 | 3 | 1 | 4 | 1 | 0.49 | 0.49 |
| 2 | −1 | 3 | −1 | 4 | 1 | 0.46 | 0.47 |
| 1 | 1 | | | 4 | −1 | 0.53 | 0.48 |
| 1 | −1 | | | 4 | −1 | 0.54 | 0.52 |
| 2 | 1 | | | 4 | −1 | 0.47 | 0.48 |
| 2 | −1 | | | 4 | −1 | 0.52 | 0.52 |
| 3 | 1 | | | 4 | −1 | 0.53 | 0.49 |
| 3 | −1 | | | 4 | −1 | 0.54 | 0.51 |
| 1 | 1 | 2 | 1 | 4 | −1 | 0.55 | 0.46 |
| 1 | 1 | 2 | −1 | 4 | −1 | 0.47 | 0.50 |
| 1 | 1 | 3 | 1 | 4 | −1 | 0.60 | 0.47 |
| 1 | 1 | 3 | −1 | 4 | −1 | 0.52 | 0.49 |
| 1 | −1 | 2 | 1 | 4 | −1 | 0.48 | 0.50 |
| 1 | −1 | 2 | −1 | 4 | −1 | 0.56 | 0.54 |
| 1 | −1 | 3 | 1 | 4 | −1 | 0.53 | 0.51 |
| 1 | −1 | 3 | −1 | 4 | −1 | 0.55 | 0.53 |
| 2 | 1 | 3 | 1 | 4 | −1 | 0.48 | 0.47 |
| 2 | 1 | 3 | −1 | 4 | −1 | 0.47 | 0.49 |
| 2 | −1 | 3 | 1 | 4 | −1 | 0.45 | 0.51 |
| 2 | −1 | 3 | −1 | 4 | −1 | 0.54 | 0.53 |

## Table A-3

*Observed and Predicted Probability Judgments for Experiments 4 and 5*

| Given | | Given | | Judged | | Experiment 4 | | Experiment 5 | |
|---|---|---|---|---|---|---|---|---|---|
| Feature | Value | Feature | Value | Feature | Value | Observed probability | 2-step model | Observed probability | 2-step model |
| 2 | 1 | | | 1 | 1 | 0.75 | 0.68 | 0.69 | 0.66 |
| 2 | −1 | | | 1 | 1 | 0.23 | 0.32 | 0.32 | 0.35 |
| 3 | 1 | | | 1 | 1 | 0.59 | 0.55 | 0.57 | 0.54 |
| 3 | −1 | | | 1 | 1 | 0.47 | 0.46 | 0.46 | 0.46 |
| 4 | 1 | | | 1 | 1 | 0.51 | 0.50 | 0.57 | 0.50 |
| 4 | −1 | | | 1 | 1 | 0.49 | 0.50 | 0.49 | 0.50 |
| 2 | 1 | 3 | 1 | 1 | 1 | 0.67 | 0.70 | 0.64 | 0.68 |
| 2 | 1 | 3 | −1 | 1 | 1 | 0.62 | 0.64 | 0.64 | 0.61 |
| 2 | 1 | 4 | 1 | 1 | 1 | 0.70 | 0.66 | 0.63 | 0.63 |
| 2 | 1 | 4 | −1 | 1 | 1 | 0.65 | 0.66 | 0.55 | 0.63 |
| 2 | −1 | 3 | 1 | 1 | 1 | 0.38 | 0.36 | 0.38 | 0.39 |
| 2 | −1 | 3 | −1 | 1 | 1 | 0.35 | 0.30 | 0.38 | 0.32 |
| 2 | −1 | 4 | 1 | 1 | 1 | 0.38 | 0.34 | 0.38 | 0.37 |
| 2 | −1 | 4 | −1 | 1 | 1 | 0.31 | 0.34 | 0.35 | 0.37 |
| 3 | 1 | 4 | 1 | 1 | 1 | 0.62 | 0.53 | 0.57 | 0.53 |
| 3 | 1 | 4 | −1 | 1 | 1 | 0.51 | 0.54 | 0.51 | 0.53 |
| 3 | −1 | 4 | 1 | 1 | 1 | 0.48 | 0.46 | 0.52 | 0.47 |
| 3 | −1 | 4 | −1 | 1 | 1 | 0.50 | 0.47 | 0.46 | 0.47 |
| 2 | 1 | | | 1 | −1 | 0.26 | 0.32 | 0.30 | 0.35 |
| 2 | −1 | | | 1 | −1 | 0.78 | 0.68 | 0.74 | 0.66 |
| 3 | 1 | | | 1 | −1 | 0.54 | 0.46 | 0.49 | 0.46 |
| 3 | −1 | | | 1 | −1 | 0.55 | 0.55 | 0.56 | 0.54 |
| 4 | 1 | | | 1 | −1 | 0.49 | 0.50 | 0.50 | 0.50 |
| 4 | −1 | | | 1 | −1 | 0.48 | 0.50 | 0.55 | 0.50 |
| 2 | 1 | 3 | 1 | 1 | −1 | 0.34 | 0.30 | 0.36 | 0.32 |
| 2 | 1 | 3 | −1 | 1 | −1 | 0.36 | 0.36 | 0.35 | 0.39 |
| 2 | 1 | 4 | 1 | 1 | −1 | 0.35 | 0.34 | 0.36 | 0.37 |
| 2 | 1 | 4 | −1 | 1 | −1 | 0.34 | 0.34 | 0.40 | 0.37 |
| 2 | −1 | 3 | 1 | 1 | −1 | 0.60 | 0.64 | 0.67 | 0.61 |
| 2 | −1 | 3 | −1 | 1 | −1 | 0.69 | 0.70 | 0.68 | 0.68 |
| 2 | −1 | 4 | 1 | 1 | −1 | 0.69 | 0.66 | 0.65 | 0.63 |
| 2 | −1 | 4 | −1 | 1 | −1 | 0.66 | 0.66 | 0.68 | 0.63 |
| 3 | 1 | 4 | 1 | 1 | 1 | 0.46 | 0.47 | 0.48 | 0.47 |
| 3 | 1 | 4 | −1 | 1 | −1 | 0.48 | 0.46 | 0.50 | 0.47 |
| 3 | −1 | 4 | 1 | 1 | −1 | 0.53 | 0.54 | 0.52 | 0.53 |
| 3 | −1 | 4 | −1 | 1 | −1 | 0.54 | 0.53 | 0.59 | 0.53 |
| 1 | 1 | | | 2 | 1 | 0.74 | 0.69 | 0.72 | 0.66 |
| 1 | −1 | | | 2 | 1 | 0.29 | 0.31 | 0.31 | 0.34 |
| 3 | 1 | | | 2 | 1 | 0.61 | 0.53 | 0.62 | 0.53 |
| 3 | −1 | | | 2 | 1 | 0.46 | 0.47 | 0.46 | 0.47 |
| 4 | 1 | | | 2 | 1 | 0.51 | 0.53 | 0.48 | 0.52 |
| 4 | −1 | | | 2 | 1 | 0.49 | 0.47 | 0.48 | 0.48 |
| 1 | 1 | 3 | 1 | 2 | 1 | 0.71 | 0.68 | 0.62 | 0.65 |
| 1 | 1 | 3 | −1 | 2 | 1 | 0.60 | 0.65 | 0.61 | 0.61 |
| 1 | 1 | 4 | 1 | 2 | 1 | 0.67 | 0.67 | 0.64 | 0.64 |
| 1 | 1 | 4 | −1 | 2 | 1 | 0.63 | 0.65 | 0.64 | 0.62 |
| 1 | −1 | 3 | 1 | 2 | 1 | 0.41 | 0.35 | 0.40 | 0.39 |
| 1 | −1 | 3 | −1 | 2 | 1 | 0.34 | 0.32 | 0.35 | 0.35 |
| 1 | −1 | 4 | 1 | 2 | 1 | 0.37 | 0.36 | 0.42 | 0.38 |
| 1 | −1 | 4 | −1 | 2 | 1 | 0.34 | 0.33 | 0.32 | 0.36 |
| 3 | 1 | 4 | 1 | 2 | 1 | 0.65 | 0.56 | 0.57 | 0.55 |
| 3 | 1 | 4 | −1 | 2 | 1 | 0.52 | 0.50 | 0.53 | 0.52 |
| 3 | −1 | 4 | 1 | 2 | 1 | 0.45 | 0.50 | 0.48 | 0.48 |
| 3 | −1 | 4 | −1 | 2 | 1 | 0.45 | 0.44 | 0.42 | 0.45 |

Table A-3 *Continued*

| Given | | Given | | Judged | | Experiment 4 | | Experiment 5 | |
|---|---|---|---|---|---|---|---|---|---|
| Feature | Value | Feature | Value | Feature | Value | Observed probability | 2-step model | Observed probability | 2-step model |
| 1 | 1 | | | 2 | −1 | 0.28 | 0.31 | 0.33 | 0.34 |
| 1 | −1 | | | 2 | −1 | 0.81 | 0.69 | 0.71 | 0.66 |
| 3 | 1 | | | 2 | −1 | 0.40 | 0.47 | 0.45 | 0.47 |
| 3 | −1 | | | 2 | −1 | 0.67 | 0.53 | 0.63 | 0.53 |
| 4 | 1 | | | 2 | −1 | 0.51 | 0.47 | 0.51 | 0.48 |
| 4 | −1 | | | 2 | −1 | 0.48 | 0.53 | 0.56 | 0.52 |
| 1 | 1 | 3 | 1 | 2 | −1 | 0.32 | 0.32 | 0.35 | 0.35 |
| 1 | 1 | 3 | −1 | 2 | −1 | 0.40 | 0.35 | 0.45 | 0.39 |
| 1 | 1 | 4 | 1 | 2 | −1 | 0.37 | 0.33 | 0.37 | 0.36 |
| 1 | 1 | 4 | −1 | 2 | −1 | 0.36 | 0.36 | 0.43 | 0.38 |
| 1 | −1 | 3 | 1 | 2 | −1 | 0.66 | 0.65 | 0.63 | 0.61 |
| 1 | −1 | 3 | −1 | 2 | −1 | 0.74 | 0.68 | 0.68 | 0.65 |
| 1 | −1 | 4 | 1 | 2 | −1 | 0.70 | 0.65 | 0.63 | 0.62 |
| 1 | −1 | 4 | −1 | 2 | −1 | 0.70 | 0.67 | 0.64 | 0.64 |
| 3 | 1 | 4 | 1 | 2 | −1 | 0.51 | 0.44 | 0.45 | 0.45 |
| 3 | 1 | 4 | −1 | 2 | −1 | 0.50 | 0.50 | 0.46 | 0.48 |
| 3 | −1 | 4 | 1 | 2 | −1 | 0.52 | 0.50 | 0.53 | 0.52 |
| 3 | −1 | 4 | −1 | 2 | −1 | 0.58 | 0.56 | 0.59 | 0.55 |
| 1 | 1 | | | 3 | 1 | 0.57 | 0.57 | 0.57 | 0.56 |
| 1 | −1 | | | 3 | 1 | 0.46 | 0.43 | 0.50 | 0.44 |
| 2 | 1 | | | 3 | 1 | 0.62 | 0.57 | 0.65 | 0.56 |
| 2 | −1 | | | 3 | 1 | 0.40 | 0.43 | 0.49 | 0.44 |
| 4 | 1 | | | 3 | 1 | 0.64 | 0.64 | 0.62 | 0.56 |
| 4 | 1 | | | 3 | 1 | 0.35 | 0.36 | 0.48 | 0.44 |
| 1 | 1 | 2 | 1 | 3 | 1 | 0.67 | 0.61 | 0.58 | 0.60 |
| 1 | 1 | 2 | −1 | 3 | 1 | 0.50 | 0.50 | 0.49 | 0.50 |
| 1 | 1 | 4 | 1 | 3 | 1 | 0.67 | 0.63 | 0.55 | 0.58 |
| 1 | 1 | 4 | −1 | 3 | 1 | 0.45 | 0.47 | 0.51 | 0.50 |
| 1 | −1 | 2 | 1 | 3 | 1 | 0.54 | 0.50 | 0.53 | 0.50 |
| 1 | −1 | 2 | −1 | 3 | 1 | 0.48 | 0.39 | 0.48 | 0.40 |
| 1 | −1 | 4 | 1 | 3 | 1 | 0.56 | 0.53 | 0.51 | 0.50 |
| 1 | −1 | 4 | −1 | 3 | 1 | 0.42 | 0.37 | 0.50 | 0.42 |
| 2 | 1 | 4 | 1 | 3 | 1 | 0.63 | 0.63 | 0.63 | 0.58 |
| 2 | 1 | 4 | −1 | 3 | 1 | 0.48 | 0.47 | 0.55 | 0.50 |
| 2 | −1 | 4 | 1 | 3 | 1 | 0.54 | 0.53 | 0.52 | 0.50 |
| 2 | −1 | 4 | −1 | 3 | 1 | 0.43 | 0.37 | 0.47 | 0.42 |
| 1 | 1 | | | 3 | −1 | 0.53 | 0.43 | 0.50 | 0.44 |
| 1 | −1 | | | 3 | −1 | 0.50 | 0.57 | 0.56 | 0.56 |
| 2 | 1 | | | 3 | −1 | 0.45 | 0.43 | 0.43 | 0.44 |
| 2 | −1 | | | 3 | −1 | 0.60 | 0.57 | 0.64 | 0.56 |
| 4 | 1 | | | 3 | −1 | 0.37 | 0.36 | 0.45 | 0.44 |
| 4 | 1 | | | 3 | −1 | 0.70 | 0.64 | 0.60 | 0.56 |
| 1 | 1 | 2 | 1 | 3 | −1 | 0.43 | 0.39 | 0.53 | 0.40 |
| 1 | 1 | 2 | −1 | 3 | −1 | 0.50 | 0.50 | 0.55 | 0.50 |
| 1 | 1 | 4 | 1 | 3 | −1 | 0.42 | 0.37 | 0.45 | 0.42 |
| 1 | 1 | 4 | −1 | 3 | −1 | 0.57 | 0.53 | 0.47 | 0.50 |
| 1 | −1 | 2 | 1 | 3 | −1 | 0.53 | 0.50 | 0.47 | 0.50 |
| 1 | −1 | 2 | −1 | 3 | −1 | 0.62 | 0.61 | 0.59 | 0.60 |
| 1 | −1 | 4 | 1 | 3 | −1 | 0.49 | 0.47 | 0.48 | 0.50 |
| 1 | −1 | 4 | −1 | 3 | −1 | 0.64 | 0.63 | 0.59 | 0.58 |
| 2 | 1 | 4 | 1 | 3 | −1 | 0.45 | 0.37 | 0.42 | 0.42 |
| 2 | 1 | 4 | −1 | 3 | −1 | 0.60 | 0.53 | 0.53 | 0.50 |
| 2 | −1 | 4 | 1 | 3 | −1 | 0.48 | 0.47 | 0.52 | 0.50 |
| 2 | −1 | 4 | −1 | 3 | −1 | 0.61 | 0.63 | 0.61 | 0.58 |
| 1 | 1 | | | 4 | 1 | 0.53 | 0.50 | 0.52 | 0.50 |
| 1 | −1 | | | 4 | 1 | 0.52 | 0.50 | 0.54 | 0.50 |

Table A-3 *Continued*

| Given | | Given | | Judged | | Experiment 4 | | Experiment 5 | |
|---|---|---|---|---|---|---|---|---|---|
| Feature | Value | Feature | Value | Feature | Value | Observed probability | 2-step model | Observed probability | 2-step model |
| 2 | 1 | | | 4 | 1 | 0.57 | 0.55 | 0.54 | 0.53 |
| 2 | −1 | | | 4 | 1 | 0.49 | 0.46 | 0.51 | 0.47 |
| 3 | 1 | | | 4 | 1 | 0.65 | 0.63 | 0.64 | 0.56 |
| 3 | −1 | | | 4 | 1 | 0.35 | 0.37 | 0.43 | 0.44 |
| 1 | 1 | 2 | 1 | 4 | 1 | 0.55 | 0.52 | 0.54 | 0.51 |
| 1 | 1 | 2 | −1 | 4 | 1 | 0.46 | 0.47 | 0.51 | 0.48 |
| 1 | 1 | 3 | 1 | 4 | 1 | 0.65 | 0.59 | 0.59 | 0.55 |
| 1 | 1 | 3 | −1 | 4 | 1 | 0.38 | 0.41 | 0.50 | 0.46 |
| 1 | −1 | 2 | 1 | 4 | 1 | 0.55 | 0.53 | 0.46 | 0.52 |
| 1 | −1 | 2 | −1 | 4 | 1 | 0.48 | 0.48 | 0.51 | 0.49 |
| 1 | −1 | 3 | 1 | 4 | 1 | 0.54 | 0.59 | 0.48 | 0.55 |
| 1 | −1 | 3 | −1 | 4 | 1 | 0.40 | 0.41 | 0.45 | 0.46 |
| 2 | 1 | 3 | 1 | 4 | 1 | 0.62 | 0.66 | 0.54 | 0.58 |
| 2 | 1 | 3 | −1 | 4 | 1 | 0.45 | 0.42 | 0.44 | 0.47 |
| 2 | −1 | 3 | 1 | 4 | 1 | 0.58 | 0.58 | 0.50 | 0.53 |
| 2 | −1 | 3 | −1 | 4 | 1 | 0.43 | 0.34 | 0.44 | 0.42 |
| 1 | 1 | | | 4 | −1 | 0.47 | 0.50 | 0.48 | 0.50 |
| 1 | −1 | | | 4 | −1 | 0.51 | 0.50 | 0.57 | 0.50 |
| 2 | 1 | | | 4 | −1 | 0.47 | 0.46 | 0.48 | 0.47 |
| 2 | −1 | | | 4 | −1 | 0.50 | 0.55 | 0.54 | 0.53 |
| 3 | 1 | | | 4 | −1 | 0.35 | 0.37 | 0.42 | 0.44 |
| 3 | −1 | | | 4 | −1 | 0.68 | 0.63 | 0.61 | 0.56 |
| 1 | 1 | 2 | 1 | 4 | −1 | 0.50 | 0.48 | 0.49 | 0.49 |
| 1 | 1 | 2 | −1 | 4 | −1 | 0.52 | 0.53 | 0.50 | 0.52 |
| 1 | 1 | 3 | 1 | 4 | −1 | 0.39 | 0.41 | 0.48 | 0.46 |
| 1 | 1 | 3 | −1 | 4 | −1 | 0.62 | 0.59 | 0.48 | 0.55 |
| 1 | −1 | 2 | 1 | 4 | −1 | 0.48 | 0.47 | 0.51 | 0.48 |
| 1 | −1 | 2 | −1 | 4 | −1 | 0.54 | 0.52 | 0.57 | 0.51 |
| 1 | −1 | 3 | 1 | 4 | −1 | 0.39 | 0.41 | 0.50 | 0.46 |
| 1 | −1 | 3 | −1 | 4 | −1 | 0.60 | 0.59 | 0.56 | 0.55 |
| 2 | 1 | 3 | 1 | 4 | −1 | 0.41 | 0.34 | 0.49 | 0.42 |
| 2 | 1 | 3 | −1 | 4 | −1 | 0.59 | 0.58 | 0.55 | 0.53 |
| 2 | −1 | 3 | 1 | 4 | −1 | 0.45 | 0.42 | 0.50 | 0.47 |
| 2 | −1 | 3 | −1 | 4 | −1 | 0.58 | 0.66 | 0.58 | 0.58 |

## REFERENCES


Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, **98**, 409–429.

Barsalou, L. W. (1990). On the indistinguishability of exemplar memory and abstraction in category representation. In T. K. Srull & R. S. Weyer (Eds.), *Advances in social cognition*. Hillsdale, NJ: Erlbaum.

Billman, D., & Heit, E. (1988). Observational learning from internal feedback: A simulation of an adaptive learning method. *Cognitive Science*, **12**, 587–625.

Brooks, L. (1978). Nonanalytic concept formation and memory for instances. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization*. Hillsdale, NJ: Erlbaum.

Chandler, P. J. (1965). *Subroutine STEPIT: An algorithm that finds the values of the parameters which minimize a given continuous function*. [Computer program]. Bloomington: Indiana University, Quantum Chemistry Program Exchange.

Fried, L. S., & Holyoak, K. J. (1984). Induction of category distributions: A framework for classification learning. *Journal of Experimental Psychology: Learning, Memory and Cognition*, **10**, 234–257.

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

Griggs, R. A. (1976). Logical processing of set inclusion relations in meaningful text. *Memory & Cognition*, **4**, 730–740.

Heit, E. (1990). *Reasoning from examples*. Doctoral dissertation, Stanford University.

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review*, **93**, 411–428.

Holland, J. H., Holyoak, K. J., Nisbett, R. E., & Thagard, P. R. (1986). *Induction: Processes of inference, learning, and discovery*. Cambridge, MA: MIT Press.

Kahneman, D., & Miller, D. T. (1986). Norm theory: Comparing reality to its alternatives. *Psychological Review*, **93**, 136–153.

Klayman, J. (1988). On the how and why (not) of learning from outcomes. In B. Brehmer & C. R. B. Joyce (Eds.), *Human judgment: The SJT view*. Amsterdam: Elsevier.

Kolodner, J. L., & Simpson, R. L. (1989). The MEDIATOR: Analysis of an early case-based problem solver. *Cognitive Science*, **13**, 507–549.

Markman, E. M. (1989). *Categorization and naming in children: Problems of induction*. Cambridge, MA: MIT Press.

McClelland, J. L. (1981). Retrieving general and specific information from stored knowledge of specifics. *Proceedings of the Third Annual Conference of the Cognitive Science Society*, 170–172.

McNamara, T. P., & Altarriba, J. (1988). Depth of spreading activation revisited: Semantic mediated priming occurs in lexical decisions. *Journal of Memory and Language*, **27**, 545–559.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, **85**, 207–238.

Mill, J. S. (1874). *A system of logic, ratiocinative and inductive*. New York: Harper Brothers.

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

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

Nosofsky, R. M. (1984). Choice, similarity and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory and Cognition*, **10**, 104–114.

Nosofsky, R. M. (1988). Similarity, frequency, and category representations. *Journal of Experimental Psychology: Learning, Memory and Cognition*, **14**, 54–65.

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

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, **83**, 304–308.

Ratcliff, R., & McKoon, G. (1988). A retrieval theory of priming in memory. *Psychological Review*, **95**, 385–408.

Riesbeck, C. K., & Schank, R. C. (1989). *Inside case-based reasoning*. Hillsdale, NJ: Erlbaum.

Rosch, E., Simpson, C., & Miller, R. S. (1976). Structural bases of typicality effects. *Journal of Experimental Psychology: Human Perception and Performance*, **2**, 491–502.

Ross, B. H., Perkins, S. J., & Tenpenny, P. L. (1990). Reminding-based category learning. *Cognitive Psychology*, **22**, 460–492.

Rumelhart, D. E., & Zipser, D. (1985). Feature discovery by competitive learning. *Cognitive Science*, **9**, 75–112.

Schneider, W. (1988). Micro Experimental Laboratory: An integrated system for IBM PC compatibles. *Behavioral Research Methods, Instruments, & Computers*, **20**, 206–217.

(Accepted August 15, 1991)