Memory & Cognition
2005, 33 (3), 479-487

Influence of multiple categories on the prediction of unknown properties

MICHAEL F. VERDE
University of Massachusetts, Amherst, Massachusetts

GREGORY L. MURPHY
New York University, New York, New York

and

BRIAN H. ROSS
University of Illinois at Urbana-Champaign, Urbana, Illinois

Knowing an item's category helps us predict its unknown properties. Previous studies suggest that when asked to evaluate the probability of an unknown property, people tend to consider only an item's most likely category, ignoring alternative categories. In the present study, property prediction took the form of either a probability rating or a speeded binary-choice judgment. In keeping with past findings, the subjects ignored alternative categories in their probability ratings. However, their binary-choice judgments were influenced by alternative categories. This novel finding suggests that the way in which category knowledge is used in prediction depends critically on the form of the prediction.

Imagine walking through the forest when a small, brown creature darts across your path. Do you keep walking or back away? Your reaction depends on whether you think an animal with the glimpsed features is likely to be dangerous. The need to predict the likelihood of an unknown property (*dangerous*) given some known properties (*small* and *brown*) is something that confronts us on a daily basis. Past experience, in the form of knowledge about categories, can be used to guide such predictions. For example, if the small, brown creature is categorized as a rodent, knowing the typical properties of rodents helps to assess the danger posed by the unknown animal. Support for inductive prediction is thought to be one of the central functions of categories (Anderson, 1991).

The use of categories in making predictions is complicated by the fact that category membership is often uncertain. The creature might be a rodent, but given its known properties it might instead be a snake or a bird. The alternative categories differ considerably in their association with the property *dangerous*, so that the decision to exercise caution depends on the category.

There are two approaches people can take when categorization is uncertain. Suppose you were to meet a person at a party who asks if you want to have coffee during the week. This person is talkative and outgoing . . . perhaps a little too outgoing. In all likelihood, these traits indicate that the person is *friendly*, in which case you would be interested in a coffee date. If, on the other hand, these traits are signs that the person is *overbearing*, you would be less interested. A prediction must be made (Will the person be an enjoyable coffee companion?) before categorization is certain. With a single-category approach, you choose the most likely category and use that to make your prediction. That is, if you decide the person is probably just friendly, then you act toward him or her as you would toward any friendly person, and the possibility that the person might be overbearing does not influence your prediction. With a multiple-category approach, on the other hand, you consider all possible categorizations when making your prediction. If there is some chance that the person is overbearing, you will be less likely to agree to have coffee than you would be if you were absolutely sure that the person is only friendly. That is, your prediction would reflect both possible categories.

The multiple-category approach conforms to a classically rational view of decision making in which people attempt to maximize accuracy by using what amounts to a naive form of statistics. Many models of inductive reasoning are based on the idea that people consider multiple categories during the decision process. However, a body of empirical evidence suggests that when predicting unknown properties as described above, people generally adopt a single-category approach. In the real world, property prediction takes many different forms, and it may be that different prediction tasks lead people to take different approaches to prediction. In the present study, we were interested in whether people use cate-

This research was supported by Grant NIMH41704 to G.L.M. M.F.V. was also supported by Training Grant NIMH16745-19. The authors are grateful to Joshua Tenenbaum and two anonymous reviewers for their helpful comments on an earlier version of the manuscript. Correspondence concerning this article should be addressed to M. F. Verde, School of Psychology, University of Plymouth, Drake Circus, Plymouth, Devon PL4 8AA, England (e-mail: michael.verde@plymouth.ac.uk).

gories differently when property prediction takes the form of specific probability ratings in comparison with speeded binary-choice judgments.

The Multiple-Category Approach

An example of the multiple-category approach can be found in Bayesian models of induction, which have been successfully applied to several types of tasks. For example, Tenenbaum (1999, 2000) examined the problem of generalizing from known examples of a category to novel examples. In one experiment (Tenenbaum, 2000), subjects were shown numbers generated according to a simple mathematical rule (e.g., powers of 10, multiples of 2, etc.). Without being told the actual rule, they had to decide whether or not a novel set of numbers also followed the same rule. The initial set could have been generated by any one of a variety of possible mathematical rules, and the task implicitly required deciding which of these rules connected the initial set of numbers to the novel set. A Bayesian model in which people consider all possible rules, weighting each by its likelihood given the initial number set, fit the data well. If we draw an analogy from these rules and number sets to categories and exemplars, we would expect people to use multiple categories in making predictions.

Heit (1998) applied a similar model to the problem of generalizing a specific property across categories. Typically, this problem is expressed as a syllogism—the premise that some categories possess a property (robins use serotonin; sparrows use serotonin) and the conclusion that some other category also has the property (penguins use serotonin). Although the true range of the property might be unknown, the universe of possibilities can be divided into a number of hypotheses (e.g., robins, sparrows, and penguins use serotonin; robins and penguins but not sparrows use serotonin). To judge the likelihood of the conclusion, one might consider all of these hypotheses, each weighted by its likelihood. Such a model can account for a number of well-known phenomena, such as the effects of the typicality and diversity of the premise categories. If we think of hypotheses as categories, this is further reason to expect people to use multiple categories in making predictions.

Anderson (1991) specifically dealt with the problem of property prediction in his Bayesian model of categorization and induction. According to the model, people consider all categories that might apply to the known features, with the influence of each category weighted by the likelihood of that category. To predict whether or not an outing with the person from the party would be enjoyable, one would consider the probability of enjoying a friendly person's company weighted by the probability of the person's being friendly, combined with the probability of enjoying an overbearing person's company weighted by the probability of the person's being overbearing.

Despite some success with phenomena related to categorization, Anderson's (1991) model has not been well supported by findings from studies of property prediction. This failure is puzzling given the success of similar models with other forms of inductive reasoning, and it suggests that we need to examine the problem of property prediction more closely before we can understand how it relates to other problems of induction.

The Single-Category Approach

Many previous studies suggest that people adopt a single-category approach to predicting unknown properties. Murphy and Ross (1994) showed subjects geometric figures drawn by 4 children, each of whom had preferences for certain shapes and colors. Thus, a given shape or color was associated most with a particular child. This child was the *dominant category* for that feature. The feature was also associated to a lesser extent with a second child, who was the *alternative category* for that feature. During a subsequent test, the subjects were told that a figure drawn by one of the four children (whose name was not disclosed) had a certain shape or color. Their task was to predict the identity of the other feature and rate its likelihood. (An example of a test question would be, "If a drawing is green, what shape is it, and how likely is it to be that shape?") Categories help prediction because we know the likelihood that a member of a category will have a given property. In the present example, knowing which child drew the figure says something about its most likely features as well as the probability of those features. For a green figure, there are two possible artists—in other words, two relevant categories (dominant and alternative). If people consider both of these categories, as Anderson's (1991) model suggests, then predictions should be affected by the base rates of features within both the dominant and alternative categories. Murphy and Ross (1994) held the rates of critical features within dominant categories constant while manipulating the rates within alternative categories. The manipulation had no effect on predictions or probability ratings. This finding suggests that the subjects considered only the dominant category and ignored the alternative category. The result was not due simply to inattention or insensitivity to categories, because the subjects were strongly affected when changes were made to the dominant category itself (see also Murphy & Ross, 2005).

Malt, Ross, and Murphy (1995) generalized this finding to real-world categories. For example, subjects read a story that described a person who seemed most likely to be a *real estate agent*. However, the story suggested a second, albeit less likely, category: either *burglar* or *cable repairman*. The subjects were then asked, "What is the probability that the man will pay attention to the sturdiness of the doors on the house?" This behavior is more consistent with the activity of a burglar than with that of a cable repairman. If predictions are influenced by the alternative category, it would be expected that those given the alternative category *burglar* would rate the behavior as being more likely than those given the category *cable repairman*. There was no difference in the likelihood ratings given by the two groups. Ross and Murphy

(1996) replicated this result but discovered that subjects could be made to consider the alternative category when the question explicitly primed that category. For example, the question "What is the probability that the man will try to find out whether Mrs. Sullivan keeps her windows locked?" describes a behavior highly associated with *burglar*. The alternative category did influence predictions when primed in this way, but the effect was short-lived and did not affect subsequent predictions. It seems that in most circumstances, uncertainty about initial categorization is not carried over from the categorization process to the prediction process. Ross and Murphy suggested that the failure to attend to alternative categories is not a limitation of the prediction process per se. People can be made to attend to multiple categories, but they generally do not do so unless an alternative category is made especially salient.

The Nature of Prediction

Consider the single-category heuristic as a strategic use of category information. It is hard to imagine a categorization process that does not involve contact with a variety of alternative categories. Malt et al. (1995) therefore concluded that the information passed on to the prediction stage must be a subset of what is available earlier in processing. This might seem odd, given that discarding information sacrifices maximum accuracy. However, in real-world situations accuracy must often be weighed against factors such as efficiency and effort (Nisbett & Ross, 1980; Simon, 1969). Taking into account all of the alternative categories may be computationally difficult or time-consuming. Moreover, in the real world an object typically belongs to many potential categories, most of which will be unlikely or not very relevant. Considering them would have little effect on accuracy. Focusing on only the most important or relevant category is a compromise that would in most cases produce acceptable levels of accuracy. With a less demanding task, such as simple-choice prediction, which does not require a numerically specific answer, there may be less of a need for the single-category heuristic.

Furthermore, consider how people approach the problem of rating probability. If something seems likely, why say that it is 65% likely rather than 55% or 75% likely? A numerically specific answer seems to call for a formal solution. The normatively correct solution would involve a Bayesian computation. However, such a computation becomes increasingly complex when one begins to consider multiple categories, which makes a single-category heuristic attractive. Simple-choice predictions, on the other hand, do not require a formal solution but rather are often made quickly and intuitively. It has often been suggested that intuitive, automatic forms of reasoning draw upon processes different than those tapped by reasoning that takes a formal or analytical approach (for a review, see Evans & Over, 1996, and Stanovich & West, 2000). Models of speeded simple-choice judgments in domains such as categorization and recognition emphasize the role of memory retrieval or associative learning (Lamberts, 2002; Logan, 1988; Nosofsky & Palmeri, 1997). Such mechanisms may operate at a precategorical level, drawing on information across categorical boundaries.

People seem to take a single-category approach when asked to predict specific probabilities—perhaps a reasonable strategy given the demands of that task. However, it is important to know whether people take the same approach to simpler types of predictions. In the present study, we focused on the empirical question of whether or not speeded simple-choice predictions are also influenced by only a single, dominant category. In Experiments 1 and 2, we used materials similar to those of Murphy and Ross (1994). In their study, in which no evidence of the use of multiple categories was found, subjects were asked to predict specific probabilities (e.g., the probability that an object has feature X given that it has feature Y). We replicated their results here. However, we found evidence of multiple-category use in response time (RT) when the task required speeded binary-choice predictions. Experiment 3 replicated this result with new materials and category structure. These novel findings stand in contrast to the consistent past failures within this paradigm to find evidence that multiple categories influence the prediction of an unknown property on the basis of known properties (Malt et al., 1995; Murphy & Ross, 1994, 2005; Ross & Murphy, 1996).

EXPERIMENT 1

How do people use category knowledge to predict an item's unknown characteristics? In Experiment 1, we returned to the artificial categories used by Murphy and Ross (1994). Pictures of colored shapes were drawn by four children: Bob, John, Sam, and Ed. Each of these four categories was defined by the particular shapes and colors preferred by each child (e.g., Bob liked to draw triangles and color his shapes green). After learning the categories, subjects were asked to make predictions about the colors or shapes of new drawings.

Two pairs of critical feature*s* were of interest: *green and triangle*, and *red and square*. The co-occurrence of the features within their respective dominant (most likely) categories was the same for both pairs. In other words, Bob was the child most likely to draw green shapes and triangles, and his drawings were 75% green and 75% triangles. John was the child most likely to draw red shapes and squares, and his drawings were 75% red and 75% squares. The co-occurrence of the critical features outside of their dominant categories was different. For example, red squares were drawn only by John, although other children drew squares and red figures. This was designated the *baseline* condition. In contrast, green triangles were sometimes drawn by children other than Bob. This was designated the *increasing* condition.

Consider the question "If you know that a drawing is green, what shape do you think it is?" According to the single-category approach, one decides on the most likely category of a green drawing (Bob) and then predicts the

most likely shape within that category (triangle). Because the critical features are identically distributed within their dominant categories, use of the single-category approach should lead to no differences in predicting *triangle* given *green* (increasing condition) or *square* given *red* (baseline condition). On the other hand, with the multiple-category approach one considers the likelihood of feature co-occurrence across all categories. Among the other categories, the likelihood of a green drawing's being a triangle is greater than that of a red drawing's being a square. This should lead to a difference in predictions between the increasing and baseline conditions.

Murphy and Ross (1994) had subjects predict the unknown feature and then rate the probability (on a scale of 0% to 100%) of the predicted feature. They found no difference in probability ratings between the increasing and baseline conditions. The present experiment differed in two ways from those of Murphy and Ross (1994). First, there were many more exemplars per category, and the subjects learned the categories via exemplar training. Second, the prediction took the form of a speeded binary choice. The subjects were shown one feature of an item (the given feature) followed by a second feature (the feature to be predicted) and had to decide whether the second feature was the one more likely to appear given the first. For example, seeing the word *green* followed by the word *triangle* required a quick yes–no response to indicate whether a green drawing was more likely to be a triangle than another shape. The dependent variables were prediction accuracy and RT rather than the specific estimates of probability used in prior studies.

Method

Subjects. Twenty-four undergraduates from the University of Illinois at Urbana-Champaign were paid for their participation.

Materials and Design. The subjects were individually tested on computers, which controlled list generation, stimulus presentation, and response recording. The stimuli consisted of 2.5 × 2.5 cm colored shapes presented in paper booklets and on the computer screen.

Four categories (BOB, JOHN, SAM, and ED) represented drawings made by four children. There were 16 exemplars per category, and each exemplar had two features: shape (*square*, *triangle*, or *circle*) and color (*red*, *green*, *blue*, or *yellow*). The number of exemplars was chosen to allow the values of the feature dimensions to be distributed independently within a category. The details of category structure are presented in Table A1. Two different forms were created to counterbalance the assignment of features to conditions. The critical feature pairs, *green–triangle* and *red–square*, were assigned to the increasing and baseline conditions, respectively, in Form 1. The opposite assignment was made in Form 2. The examples discussed below describe Form 1.

BOB was the dominant category for the features *green* and *triangle*, which were most likely (60% of the time) to have been drawn by Bob. JOHN was the dominant category for the features *red* and *square*, which were most likely (60% of the time) to have been drawn by John. For each of the critical features, a drawing within its dominant category was 75% likely to have that feature. In other words, within the dominant categories of the critical features, the probability that a red drawing would be a square (and vice versa) was 75%, and the probability that a green drawing would be a triangle (and vice versa) was 75%. Thus, the distribution of critical features within their dominant categories was identical for the increasing and baseline conditions. The conditions differed, however, in the distribution of the critical features across all categories. For the baseline condition, the probability of a red drawing's being a square (and vice versa) was 45%. For the increasing condition, the probability of a green drawing's being a triangle (and vice versa) was 55%.

The subjects learned the categories from a four-page stimulus booklet. Each page showed a single category, with the category labels (BOB, JOHN, SAM, and ED) at the top and the 16 exemplars arrayed below it. Three different versions of the stimulus booklet were created for each of the two forms, each version featuring a different random ordering of the exemplars on the page. For the final test, four versions of the test list (two per form) were created that differed in the ordering of the test probes. Each of the critical features appeared four times as the given feature and four times as the unknown feature in the 56-item test list, with repetitions spaced at roughly equal intervals. The order of presentation of the critical features was counterbalanced within the list and also across different versions of the list. The remaining items contained filler features, each of which appeared roughly the same number of times. The number of trials for which the correct answer was "yes" roughly equaled the number for which the correct answer was "no." The 9-item practice test list contained only filler items.

Procedure. The 50-min session consisted of five phases: (1) a 10-min study period, (2) a training test, (3) a 5-min study period, (4) a training test, and (5) a final test. The experimenter provided brief verbal instructions at the beginning of each phase. More detailed instructions followed on the computer screen.

During each study period (Phases 1 and 3), the subjects were instructed to familiarize themselves with the types of pictures drawn by each child. Each subject studied the stimulus booklet at his or her own pace, and the experimenter removed the booklets at the end of each study period.

During each 64-trial training test (Phases 2 and 4), the exemplars from the stimulus booklet were presented in random order on the computer screen. On each trial, an empty box appeared in the center of the screen with the four category choices (BOB, JOHN, SAM, and ED), numbered 1–4, respectively, below it. After 1,000 msec, a colored shape appeared in the empty box. The subjects indicated the category to which the exemplar belonged using the keys "z," "x," "." and "/" (labeled with the numbers 1–4, respectively). They were warned that a particular color–shape combination might have appeared in more than one category, in which case they should give the best answer. This procedure emphasized to the subjects the fact that some feature combinations were associated not only with a dominant category but with an alternative category as well. Following a response, the three incorrect category labels were removed, leaving only the correct category label, the exemplar, and accuracy feedback. This display remained on the screen for 2,000 msec following a correct response and for 4,000 msec following an incorrect response.

The final test consisted of 9 practice trials followed by 56 test trials. At the beginning of each trial, a fixation line of plus signs appeared in the center of the screen for 1,000 msec. This was replaced by a feature word (*square*, *triangle*, *circle*, *red*, *green*, *blue*, or *yellow*). After a 1,000-msec delay, a second feature word (a color if the first had been a shape, or a shape if the first had been a color) appeared below the first word. On-screen prompts to the left of the display area served as reminders of the task. "A picture that is:" appeared 3 cm to the left of the location where the first feature had appeared, while "is most likely to be:" appeared 3 cm to the left of the location where the second feature had appeared. The subjects indicated whether or not the second feature was the one most likely to occur with the first (e.g., whether or not a triangle is most likely to be green). They were instructed to respond as soon as the correct answer came to them and to keep their fingers resting on the response keys at all times. Thus, although there was no response deadline, the subjects were encouraged to respond quickly. "Yes"

and "No" labels were affixed to the "1" and "3" keys of the keyboard's numeric keypad, respectively. The feature words remained on the screen until a response was made. A 2,000-msec blank screen followed each trial.

Results and Discussion

Only items containing critical feature pairs (i.e., *green–triangle*, *red–square*) were included in the analysis of the final prediction test. Accuracy and RTs were analyzed separately. Accurate responses were those in which the subjects correctly identified one member of a critical pair as the feature most likely to appear with the other member. There was no significant difference in accuracy between the baseline (94%) and increasing (92%) conditions [$t(23) = 0.68$]. However, the fact that accuracy was extremely high (a third of the subjects had 100% accuracy in both conditions) suggests that effects are more likely to be found in RT. In fact, RT for correct responses was significantly shorter in the increasing (1,280 msec) than in the baseline (1,499 msec) condition [$t(23) = 2.46$, $p < .05$].[1]

Because the two conditions had identically structured dominant categories, the faster responses in the increasing condition reflected the use of information outside of the dominant category. This is in direct contrast with the null findings of Murphy and Ross (1994, 2005), who used similar materials but a probability rating task. They found that subjects seemed to attend only to the dominant category. Note that the probability that a critical feature would appear with its dominant category was identical for both conditions, ruling out speed of categorization as the locus of the effect. The purpose of Experiment 2 was to verify that differences in the form of the prediction task, rather than other procedural differences, produced the difference in these findings.

EXPERIMENT 2

The finding that people use multiple categories when predicting an unknown property is surprising, given that past work in which similar materials were used consistently failed to find such evidence. The experiments of Murphy and Ross (1994) differed from our Experiment 1 in how subjects were exposed to the categories as well as in the form of the prediction test. Murphy and Ross (1994) showed subjects a small, representative set of exemplars, which the subjects either memorized or were allowed to consult while making decisions. In Experiment 1, on the other hand, subjects learned the feature distributions via training with a larger number of exemplars. The purpose of Experiment 2 was to rule out differences in category structure and learning procedure as the source of the novel finding in Experiment 1.

Experiment 2 was identical in most respects to Experiment 1. However, the final prediction task was that used by Murphy and Ross (1994). The subjects rated the probability (on a scale of 0% to 100%) that one member of a critical feature pair would appear with the other member of the pair. This prediction was similar in essence to that required in the task of Experiment 1. If the influence of multiple categories in Experiment 1 was due to the categories or to the learning procedure, one would expect the same results when subjects made explicit numerical predictions in Experiment 2. If multiple- category use was due to the nature of the simple-choice judgments, as we have suggested, then there should be no such effect here, in accordance with the findings of the literature.

Method

Subjects. Twenty-six undergraduates from the University of Illinois at Urbana-Champaign were paid for their participation.

Materials and Design. The stimulus materials and training booklets were identical to those used in Experiment 1. The test booklet for the prediction task contained six sets of questions. Each set contained four questions in the following form:

Q1. I have a drawing of a [feature]. What child do you think drew it?

Q2. What is the probability that the child you just named drew it?

Q3. What color/shape do you think this [feature] is?

Q4. What is the probability that this figure is the color/shape you just named?

For each set, [feature] was replaced with one of the four critical features (*triangle*, *green figure*, *square*, or *red figure*) or a filler (*circle* or *blue figure*). On each page of the test booklet, one question represented the increasing condition, one represented the baseline condition, and one referred to filler features. The order of the three types of question sets was counterbalanced across pages.

Procedure. The procedure was identical to that of Experiment 1 save for the final prediction test. The subjects were instructed to answer each question with a probability rating ranging from 0% to 100%.

Results and Discussion

Only the question sets pertaining to the critical features were analyzed. Although each question was examined separately, the four questions in a single set constituted a judgment essentially like that of Experiment 1: prediction of an unknown feature given the presence of a known feature. Questions 1–3 were used to verify that the subjects had accurate knowledge of how the critical features were distributed with respect to one another and within their dominant categories. The likelihood of correctly choosing the dominant category for the known feature (Question 1) did not differ reliably between the baseline (92%) and increasing (88%) conditions [$t(25) = 0.81$]. Only items for which the correct dominant category was chosen were included in the analysis of the remaining questions.[2] The rated probability that the known feature belonged to the dominant category (Question 2) did not differ reliably between the baseline (60%) and increasing (57%) conditions [$t(24) = 0.77$]. The data from Questions 1 and 2 suggest that the extent to which the subjects learned the category membership of critical features was the same for both conditions. Finally, the likelihood of correctly choosing the other critical feature (Question 3) did not differ reliably between the baseline (86%) and increasing (84%) conditions [$t(24) = 0.33$]. Question 4 (*What is the probability that this figure is the color/shape you just named?*) was the critical measure of prediction. The probability rating in Question 4 did

not differ reliably between the baseline (62%) and increasing (60%) conditions [$t(21) = 0.33$].[3]

Experiment 2 replicated the findings of Murphy and Ross (1994), who observed no difference between the increasing and baseline conditions. Note that in this experiment the subjects made an initial categorization judgment, which may have brought the dominant category into focus for them. Murphy and Ross (1994) examined this possibility but found that removing all reference to category (i.e., Questions 1 and 2) did not affect the property prediction (see also Murphy & Ross, 2005). Question 2, however, was useful for evaluating the observed null effect. Average probability ratings were close to 60% for this question, which is the actual proportion encountered during study. This shows that the subjects were well calibrated and quite aware that the given feature was associated with alternative categories. In fact, it can be argued that rather than focusing the subjects' attention on the dominant category, Question 2 emphasizes the availability of alternative categories, maximizing the chance of finding a difference between the increasing and baseline conditions. In sum, the results of Experiment 2 indicate that the form of the prediction task and the choice of dependent measure, rather than differences in category presentation, were critical to finding an influence of multiple categories in Experiment 1.

EXPERIMENT 3

The finding that people use categories differently depending on the form of the prediction task has important implications for the study of inductive judgments. Because previous evidence of single-category use has been so consistent (Malt et al., 1995; Murphy & Ross, 1994, 2005; Ross & Murphy, 1996), we attempted to replicate the findings of Experiment 1 with new materials. In Experiment 3, we changed the stimuli and categories and strengthened the manipulation so that there was a larger difference between the baseline and increasing conditions in how the critical traits occurred together across all categories. This was made possible by sacrificing the constraint that features vary independently within categories (as in Experiments 1 and 2), although critical features were still constrained in this way within their respective dominant categories. In addition to accuracy and RT, ratings of prediction confidence were added as a dependent measure.

The subjects were told to imagine that they were biologists examining wildlife taken from an alien planet. Categories represented the habitats on the planet, and exemplars were alien creatures sampled from these habitats. Each creature was described in terms of two feature dimensions. For example, a creature sampled from the *forest* habitat might have the features *spines* and *claws*. The design was otherwise similar to that of Experiment 1. Two of the categories were assigned to the increasing condition and the other two to the baseline condition. For each category, there was a pair of critical features which were predominantly associated with that category but which also appeared, to a lesser extent, in another category. The co-occurrence of the paired features within their respective dominant categories was the same for both increasing and baseline conditions. Outside of the dominant category, however, the critical features of the increasing condition sometimes appeared together, but those of the baseline condition never did. If people consider both the dominant and alternative categories, there should be a prediction advantage for the increasing condition relative to the baseline condition.

Method

Subjects. Thirty-eight undergraduates from the University of Massachusetts at Amherst participated for course credit.

Materials and Design. The subjects were individually tested on computers, which controlled list generation, stimulus presentation, and response recording. Assignment of category and exemplar labels to conditions and ordering of items within training and test lists was uniquely randomized for each subject.

Four categories represented different habitats (*forest*, *desert*, *swamp*, and *prairie*) on an alien planet. The 20 exemplars in each category represented creatures sampled from the habitat. Each exemplar had two features, drawn from Dimension 1 (*scales*, *hair*, *feathers*, *shell*, *fur*, and *spines*) and Dimension 2 (*tentacle*, *claws*, *teeth*, *tail*, *antenna*, and *horn*). Category details are presented in Table A2.

The increasing condition comprised Categories 1 and 2, whereas the baseline condition comprised Categories 3 and 4. There were two critical feature pairs within each category: A1 and B1 in Category 1, A2 and B2 in Category 2, A3 and B3 in Category 3, and A4 and B4 in Category 4. Each of these features appeared in its dominant category 75% of the time. Within its dominant category, a creature was 60% likely to have a particular critical feature, and a creature with one critical feature was 75% likely to have the other. Thus, the distributions of critical features within their respective dominant categories were identical for the increasing and baseline conditions. The conditions differed, however, in the distribution of the critical features across all categories. Overall, the critical features were 56% likely to appear together in the baseline condition but 81% likely to appear together in the increasing condition.

The list used in the final prediction test consisted of three 24-trial blocks. Within a block, 8 trials were used to test the critical feature pairs. Each critical feature appeared once as the given feature and once as the unknown feature. Within these trials, a critical feature was always paired with the other critical feature from the same category, so that the correct response was always "yes." Each critical feature also appeared in a lure trial, to which the correct answer was always "no." The remaining trials contained only filler features. Item order was randomized for each block.

Procedure. The 50-min session comprised five phases: (1) study, (2) a training test, (3) study, (4) a training test, and (5) a final test. During the 80-trial study phase, all exemplars were shown once. The study list was blocked by category, resulting in four blocks of 20 exemplars each. The order of categories within a list and that of exemplars within a block were randomized for each study phase. At the beginning of each trial, a fixation line of plus signs appeared in the center of the screen for 500 msec. The exemplar display, shown for 4,000 msec, consisted of four lines: the habitat, the specimen number (a unique eight-digit tag meant to differentiate the exemplars), Feature 1, and Feature 2. Each trial concluded with a 500-msec blank screen.

During the 80-trial training test phase, all exemplars were tested once in random order. At the beginning of each trial, the title "Unknown Sample" appeared at the top of the display, followed by Feature 1 and Feature 2 from a particular exemplar. Below these, the prompt "Which Habitat?" was followed by names of the four habi-

tats, numbered 1–4, and the prompt "Enter number (1–4)." Each test item represented a specific exemplar from the studied set. Thus, there was only one correct answer on a given trial (an exemplar with those features could have appeared in another category as well; the subjects were instructed simply to give the best answer). The subjects responded using the keys "1" to "4" and were given a 500-msec feedback message of "Correct!" or "Incorrect! Try Again." The trial ended only when the correct category was chosen. Each trial concluded with a 1,000-msec blank screen.

The final phase consisted of 8 practice trials followed by a 72-trial test. At the beginning of each trial, a fixation line of plus signs appeared in the center of the screen and was immediately replaced by the prompt "A creature with:" with a blank space to the right of it. After a 500-msec interval, the given feature appeared in the blank space beside the prompt. At the same time, directly below the first prompt appeared a second prompt ("Most likely has:") with a blank space to the right of it. After a 500-msec interval, the to-be-predicted feature appeared in the blank space beside the second prompt. The task was to decide whether or not the to-be-predicted feature was the one most likely to appear with the given feature and then to make a yes–no judgment using the "z" or "/" key, respectively. The subjects were instructed to make their judgments as quickly as the answer came to them and to keep their fingers resting on the keys at all times during the test. Following each yes–no judgment, the subjects indicated their confidence in the correctness of their answer on a four-point scale on which 1 = *not confident* and 4 = *very confident*. Each trial concluded with a 1,000-msec blank screen.

Results and Discussion

Only items containing critical feature pairs were included in the analysis of the final prediction test. Accuracy, RTs, and confidence were analyzed separately. Accurate responses were those in which the subjects correctly identified one member of a critical pair as the feature most likely to appear with the other member. Accuracy was significantly greater in the increasing (90%) than in the baseline (80%) condition [$t(37) = 2.38$, $p < .05$]. RTs for accurate responses were significantly shorter in the increasing (1,329 msec) than in the baseline (1,541 msec) condition [$t(37) = 2.34$, $p < .05$]. Confidence was significantly greater in the increasing (3.52) than in the baseline (3.23) condition [$t(37) = 2.36$, $p < .05$].

In all three dependent measures, there was evidence of the influence of the alternative category on prediction. When all categories were considered, the likelihood that the critical features would appear together was greater for the increasing condition than for the baseline condition, and this led to faster, more accurate, and more confident predictions. Why, unlike in Experiment 1, was there an effect on accuracy as well as on RT? Perhaps this was the result of strengthening the manipulation. Enlarging the difference between the increasing and baseline conditions may have magnified the effect of this difference. Also, the greater number of exemplars and feature dimensions seemed to reduce accuracy, which prevented what may have been a ceiling effect in Experiment 1.

GENERAL DISCUSSION

An important function of categories is to help us make predictions about something when we know only a few of its properties. Previous studies suggest that people often make predictions about unknown properties by choosing the most likely category given the known properties, and then using information about that category alone (Malt et al., 1995; Murphy & Ross, 1994, 2005; Ross & Murphy, 1996). These studies, however, focused on a particular form of prediction: People were asked to rate the specific probability of unknown properties. The present findings show that when people simply make speeded binary-choice predictions with the same materials, they do seem to be influenced by information from multiple categories. This is a novel result, which suggests that how people use category information when making predictions depends critically on the form of the prediction.

That task demands determine how information is used is perhaps not surprising. The challenge remains to identify why and under what conditions people adopt different approaches to property prediction. In everyday life, prediction commonly takes the form of a simple choice or preference: Is the animal dangerous? Will that person be pleasant company? The focus of the present study was motivated in part by the failure of previous studies to examine this important facet of property prediction. However, we also suspected that this simple form of prediction is more likely to be influenced by information from multiple categories because the probability rating task may specifically encourage the use of a single-category heuristic. Producing numerical probabilities is relatively difficult, and it seems to demand a formal rather than an intuitive solution.

The use of heuristics is typically related to the difficulty or complexity of a task. Probability rating is more difficult than simple-choice prediction because it requires a much more specific answer. There may also be an essential difference in the way people approach each task. Why rate the probability of an unknown property at 65% rather than at some other specific value? The choice requires justification—a correct answer seems to call for a formal solution or strategy. Predictions that require only a preference or simple choice, on the other hand, seem often to be based on a feeling that is intuitive or automatic. Theorists have often argued that intuitive and formal approaches to reasoning and decision making rely on different processes. Evans and Over (1996), for example, distinguish implicit and explicit processes, noting that people can be quite accurate at detecting contingencies in implicit learning tasks while being prone to fallacies and biases in statistical reasoning. Others distinguish processes that are automatic, intuitive, and based on acquired experience or associations from processes that are strategic, rational or analytic, and based on formally acquired rules (Sloman, 1996; Stanovich & West, 2000).

The present findings allow us only to speculate about the processes that underlie formal and intuitive approaches to property prediction. Perhaps subjects approach the problem of rating specific probabilities as they would any other mathematical problem: They apply a formal computation

such as calculation of the proportion of items with a given property. The explicit use of category knowledge allows them to limit the complexity of the computation. It is possible, of course, to take the same approach with binary-choice predictions, calculating a graded probability and turning it into a yes–no response. However, this leaves the question of why there is an apparent influence of multiple categories in one task and not in the other.

Perhaps speeded binary-choice predictions do not involve explicit categorization judgments at all but are based on global associations between features. Models of speeded simple-choice categorization and recognition often invoke direct memory retrieval as the basis for such judgments (Lamberts, 2002; Logan, 1988; Nosofsky & Palmeri, 1997). Applied to the property prediction task, such a model might suggest that the known property acts as a retrieval cue, bringing to mind exemplars that possess that property. In our experiments, because the critical known and unknown features were more consistently associated in the increasing condition relative to the baseline condition, evidence of the unknown feature would accumulate more quickly and more consistently in the increasing condition. This would explain the faster and more accurate predictions we observed in the increasing condition.

Although we have suggested that the subjects considered multiple categories within the same judgment, there is an alternative possibility. The subjects may have considered only a single category on a given trial but varied over trials depending on whether the dominant category or an alternative category was used. Although this interpretation is consistent with the claim that probability rating and simple-choice predictions are influenced differently by category information, it claims that both tasks rely on single categories for each prediction. Perhaps, as has just been discussed, the known property acts as a retrieval cue, bringing to mind not all exemplars with that property but just those that are also activated by virtue of having the same category membership. Although this is a possibility, we believe that the data presented here argue against it. For example, in Experiment 2 the subjects claimed that the given feature was 60% likely to appear in the dominant category—the only one in which the unknown and given features were very likely to appear together. If the subjects in Experiment 1, with the same materials and training, chose the dominant category only 60% of the time, their likelihood of predicting the unknown feature should have been no higher than 60%, when in fact it was about 93%. In sum, this interpretation of multiple-category use seems unlikely.

Earlier, we noted that some models of induction suggest that people consider multiple categories when generalizing a category to novel examples (Tenenbaum, 1999, 2000) and when generalizing a property across categories (Heit, 1998). Predicting unknown properties is in some ways a related problem, which made the frequent failure to observe multiple-category use puzzling. The present findings show that the use of categories depends critically on the specific form of the prediction task. Examining this issue with regard to other forms of inductive prediction may be necessary before we can understand how they relate to property prediction.

REFERENCES

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

Evans, J. St. B. T., & Over, D. E. (1996). *Rationality and reasoning*. Hove, U.K.: Psychology Press.

Heit, E. (1998). A Bayesian analysis of some forms of inductive reasoning. In M. Oaksford & N. Chater (Eds.), *Rational models of cognition* (pp. 248-274). New York: Oxford University Press.

Lamberts, K. (2002). Feature sampling in categorization and recognition of objects. *Quarterly Journal of Experimental Psychology*, **55A**, 141-154.

Logan, G. D. (1988). Toward an instance theory of automaticity. *Psychological Review*, **95**, 492-527.

Malt, B. C., Ross, B. H., & Murphy, G. L. (1995). Predicting features for members of natural categories when categorization is uncertain. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **21**, 646-661.

Murphy, G. L., & Ross, B. H. (1994). Predictions from uncertain categorizations. *Cognitive Psychology*, **27**, 148-193.

Murphy, G. L., & Ross, B. H. (2005). The two faces of typicality in category-based induction. *Cognition*, **95**, 175-200.

Nisbett, R. E., & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment*. Englewood Cliffs, NJ: Prentice-Hall.

Nosofsky, R. M., & Palmeri, T. J. (1997). An exemplar-based random walk model of speeded classification. *Psychological Review*, **104**, 266-300.

Ross, B. H., & Murphy, G. L. (1996). Category-based predictions: Influence of uncertainty and feature associations. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **22**, 736-753.

Simon, H. A. (1969). *The sciences of the artificial*. Cambridge, MA: MIT Press.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, **119**, 3-22.

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate. *Behavioral & Brain Sciences*, **23**, 645-726.

Tenenbaum, J. B. (1999). Bayesian modeling of human concept learning. *Advances in Neural Information Processing Systems*, **11**, 59-68.

Tenenbaum, J. B. (2000). Rules and similarity in concept learning. *Advances in Neural Information Processing Systems*, **12**, 59-65.

NOTES

1. To mitigate the influence of outliers, cutoff values two *SD*s above and below the mean were calculated for each subject. RTs above or below those values (6% of scores) were replaced with the cutoff values. This correction was applied in Experiment 3 as well (replacing 5% of scores). Analysis of uncorrected RTs did not affect the conclusions of either experiment.

2. Subjects who did not correctly identify the dominant category in Question 1 could not contribute usable values in Questions 2 and 3. Those who did not correctly identify the most likely shape/color in Question 3 could not contribute usable values in Question 4. For these reasons, 1 subject was excluded from the analysis of Questions 2 and 3, and 4 subjects were excluded from the analysis of Question 4.

3. The inclusion of trials with incorrect categorization (Question 1) did not change the conclusion: The baseline (61%) and increasing (60%) conditions did not differ reliably [$t(21) = 0.28$].

APPENDIX

Table A1
Experiments 1 and 2: Category Structure, Form 1

	Bob		John		Sam		Ed	
Exemplar	Color	Shape	Color	Shape	Color	Shape	Color	Shape
1	A1	B1	A2	B2	A2	B3	A4	B3
2	A1	B1	A2	B2	A2	B3	A4	B3
3	A1	B1	A2	B2	A2	B3	A4	B3
4	A1	B1	A2	B2	A2	B3	A4	B3
5	A1	B1	A2	B2	A2	B3	A4	B3
6	A1	B1	A2	B2	A2	B3	A4	B3
7	A1	B1	A2	B2	A1	B3	A4	B3
8	A1	B1	A2	B2	A1	B3	A4	B3
9	A1	B1	A2	B2	A1	B3	A4	B3
10	A1	B2	A2	B1	A3	B3	A4	B2
11	A1	B2	A2	B1	A3	B3	A4	B2
12	A1	B2	A2	B1	A3	B3	A4	B2
13	A3	B1	A1	B2	A2	B1	A3	B3
14	A3	B1	A1	B2	A2	B1	A3	B3
15	A3	B1	A1	B2	A1	B1	A3	B3
16	A3	B2	A1	B1	A3	B1	A3	B2

Note—There were 16 exemplars per category. Each exemplar had two features: color (A1 = *green*, A2 = *red*, A3 = *yellow*, or A4 = *blue*) and shape (B1 = *triangle*, B2 = *square*, or B3 = *circle*). In Experiments 1 and 2, there were two counterbalancing forms: Form 1 (shown here) and Form 2. The difference between the forms was in which pair of critical features (e.g., *green–triangle*, *red–square*) and which category (e.g., Bob, John) was assigned to the increasing condition (i.e., Bob) and which to the baseline condition (i.e., John).

Table A2
Experiment 3: Category Structure

	Category 1		Category 2		Category 3		Category 4	
Exemplar	Feature 1	Feature 2	Feature 1	Feature 2	Feature 1	Feature 2	Feature 1	Feature 2
1	A1	B1	A2	B2	A3	B3	A4	B4
2	A1	B1	A2	B2	A3	B3	A4	B4
3	A1	B1	A2	B2	A3	B3	A4	B4
4	A1	B1	A2	B2	A3	B3	A4	B4
5	A1	B1	A2	B2	A3	B3	A4	B4
6	A1	B1	A2	B2	A3	B3	A4	B4
7	A1	B1	A2	B2	A3	B3	A4	B4
8	A1	B1	A2	B2	A3	B3	A4	B4
9	A1	B1	A2	B2	A3	B3	A4	B4
10	A1	B5	A2	B5	A3	B6	A4	B6
11	A1	B5	A2	B5	A3	B6	A4	B6
12	A1	B5	A2	B5	A3	B6	A4	B6
13	A5	B1	A6	B2	A6	B3	A5	B4
14	A5	B1	A6	B2	A6	B3	A5	B4
15	A5	B1	A6	B2	A6	B3	A5	B4
16	A5	B5	A6	B5	A6	B6	A5	B6
17	A3	B4	A4	B3	A1	B1	A2	B2
18	A3	B4	A4	B3	A1	B1	A2	B2
19	A3	B4	A4	B3	A1	B1	A2	B2
20	A3	B4	A4	B3	A1	B1	A2	B2

Note—There were 20 exemplars per category. Category and label names were randomly assigned for each subject. Category names: *forest*, *desert*, *swamp*, *prairie*. Feature names: Dimension 1. *scales*, *hair*, *feathers*, *fur*, *spines*, *shell*. Dimension 2. *tentacle*, *claws*, *teeth*, *tail*, *antenna*, *horn*.

(Manuscript received October 14, 2003;
revision accepted for publication June 29, 2004.)