

Uncertainty in Category-Based Induction: When Do People Integrate Across Categories?

Gregory L. Murphy
New York University

Brian H. Ross
University of Illinois at Urbana-Champaign

Two experiments investigated how people perform category-based induction for items that have uncertain categorization. Whereas normative considerations suggest that people should consider multiple relevant categories, much past research has argued that people focus on only the most likely category. A new method is introduced in which responses on individual trials can be classified as using single or multiple categories, an improvement on past methods that relied on null effects as evidence for single-category use. Experiment 1 found that people did use multiple categories when the most likely category gave an ambiguous induction but that few people did so when it gave an unambiguous induction. Experiment 2 suggested that the reluctance to use multiple categories arose from a cognitive shortcut, in which only one source of information is consulted. The experiments revealed significant individual differences, suggesting that use of multiple categories is one of a number of strategies that can be used rather than being the basis for most category-based induction.

Keywords: concepts, categorization, reasoning, category-based induction, uncertainty

Much of cognition and perception can be characterized as an inferential process, in which people must decide what they are seeing or hearing, what an actor's intention is, or what future event will occur. People make such predictions based on perceptual input combined with past experience to infer the desired information. Within this conceptualization of cognition, Bayes' theorem offers a natural framework within which to understand processing, providing a means by which present evidence can be evaluated in the light of prior knowledge. Indeed, Bayesian approaches have now become popular in areas ranging from perception (Kersten, Mamassian, & Yuille, 2004) to higher-level cognition (Tenenbaum & Griffiths, 2001).

One issue in higher-level cognition that has been given a simple Bayesian analysis is that of category-based induction. People often make predictions about the unknown features of objects, such as internal properties or a future behavior. They wonder if some fruit hanging on a branch is edible or if a dog is likely to bite. One source of information about such questions is the object's category (or, as we shall see, categories). Indeed, this inductive process has been proposed as the main function that categories serve. After all, if categories could only tell us about features that we already perceive in the entity, they would not be providing any new information.

An interesting question arises when the object's category is not certain: It is some kind of small mammal, a citrus fruit of some

sort, or someone in a health profession. In these cases, one might need to access information about multiple categories in order to make the induction. For example, imagine that you have a medical condition that your doctor thinks is most likely an infection but may also possibly be a virus. You would hope that your doctor would take into account both possibilities in recommending treatment or in predicting your prognosis. For example, the doctor should not prescribe a treatment for the infection that would make you sicker if you actually have the virus.

The goal of the present study is to evaluate whether people consider multiple categories when making inductions about objects whose categorization is uncertain. We describe a new paradigm that improves on past studies by allowing us to identify individual subjects and even individual responses as using single or multiple categories. This method allows us to consider whether people differ in their induction processes and also whether manipulations that influence induction affect the population as a whole or shift only a subset of participants. We then use this new method to investigate why many people focus on a single category in making inductions.

Single Versus Multiple Categories in Induction

How should people generally take into account category knowledge when making inductions about an entity with an uncertain categorization? Anderson (1991) made a well-known proposal, which he identified with Bayesian inference. He suggested that when people are uncertain as to whether an object is in one of a number of categories, they calculate the induction (e.g., probability of an animal biting) from each category and then add these together, weighted by the likelihood of the item being in each category. For example, if you think that cats are 25% likely to bite, squirrels are 100% likely to bite, and dogs are 50% likely to bite, then for an animal that has scurried under your car and might be any of these things, you would take the weighted average of these probabilities. If the animal is 70% likely to be a cat, 20% likely to

Gregory L. Murphy, Department of Psychology, New York University, and Brian H. Ross, Department of Psychology, University of Illinois at Urbana-Champaign.

This research was supported by National Institute of Mental Health Grant MH41704. We are grateful to Harlan Harris and Barbara Malt for helpful comments on a draft of this article.

Correspondence concerning this article should be addressed to Gregory L. Murphy, Department of Psychology, New York University, 6 Washington Place, 8th Floor, New York, NY 10003. E-mail: gregory.murphy@nyu.edu

be a squirrel, and 10% likely to be a dog, then the chance of its biting you is $(.70 \times .25) + (.20 \times 1.0) + (.10 \times .50) = .425$. This procedure is normative, because it uses all of the information available, weighted by its likelihood of being correct. It is an example of *bet-hedging*, in that you did not simply focus on the most likely category, cats, and then make a prediction. If you had done so, you would have obtained a smaller probability of .25, because the high-biting dogs and squirrels were not taken into account.

Anderson (1991) called his approach a Bayesian model, and it shares with other Bayesian models the important feature that it integrates across different prior possibilities when making a prediction (e.g., Heit, 2000; Tenenbaum & Griffiths, 2001). However, Anderson’s proposal is only one of many possible Bayesian analyses of this task. For example, it assumes that the to-be-predicted feature is independent of the observed features, but other approaches might not assume this; Anderson assumes (as do we) that categories are central to the induction process, whereas a different analysis could ignore categories or use them as just one piece of information. Therefore, we describe the issue under investigation as a question of whether multiple categories are used in induction, rather than “Are people Bayesian?” Since Anderson’s approach is the best known Bayesian analysis of this task, the results clearly speak to more general issues of whether or in what way people are Bayesian, even if they do not address every possible Bayesian analysis of induction.

Of course, people may not be as normative in their judgments as Anderson’s (1991) model proposes, and we have conducted a number of experiments (reviewed below) in which people did not hedge their bets when making inductions. Given the success of Bayesian approaches to cognition and to category-based induction in particular, such findings are surprising. In past research, we have found that some task manipulations make people more or less likely to use multiple categories (Ross & Murphy, 1996), which suggests the possibility that the issue should not be phrased as “Do people use multiple categories?” but instead as, “When do people use multiple categories and when do they use only one?” The present article introduces a new method that allows us to answer this question more effectively.

Past Research

In one kind of experiment (Malt, Ross, & Murphy, 1995; Ross & Murphy, 1996), people read scenarios in which a person or an object was somewhat ambiguous. For example, a character in the story, Mrs. Sullivan, was expecting a visit from a real estate agent. However, the story also mentioned another possibility: either (depending on the condition) a worker for a cable TV company or a burglar who was breaking into houses in the area. At the end of the story, someone walked up the driveway of the house, and the person was not identified, though it was indicated that Ms. Sullivan thought that the person was most likely the real estate agent. Thus, real estate agent was the *target category*, and burglar or cable worker, whichever was mentioned for a given story, was the *alternative category*.

After reading the story, subjects made predictions about what the person in the story might do. The questions were designed so that some of them would have higher probabilities for one of the alternative categories, and others would have higher probabilities for the other alternative category. For example, the question,

“What is the probability that the man who walked up the driveway will ring the doorbell?” should have a higher answer for cable TV workers than for burglars. Therefore, just as in our example of an animal scurrying under the car, if people are thinking, “A real estate agent might ring the doorbell, and so would a cable TV worker about to dig in the yard,” then they should give a higher probability than if they were given the other alternative category (“A real estate agent might ring the doorbell, but a burglar definitely wouldn’t”). However, we have consistently failed to find this effect (Malt et al., 1995; Ross & Murphy, 1996). For those subjects who think that the person is most likely the real estate agent, it makes no difference what the other category is—they do not seem to hedge their bets in the way that the Anderson rule requires. Instead, they seem to be using a *single-category strategy*. That is, they think, “the person is most likely the real estate agent, and real estate agents . . .” without thinking about other categories.

Because it is difficult to directly manipulate the properties or categorization probabilities of familiar concepts such as burglars, we have also used a different methodology that provides greater experimental control over the materials. In this method, people study a visual display containing a set of objects: schematic faces, geometric forms, or drawings of objects. The objects are divided into categories (see Figure 1), either specifying the identity of the

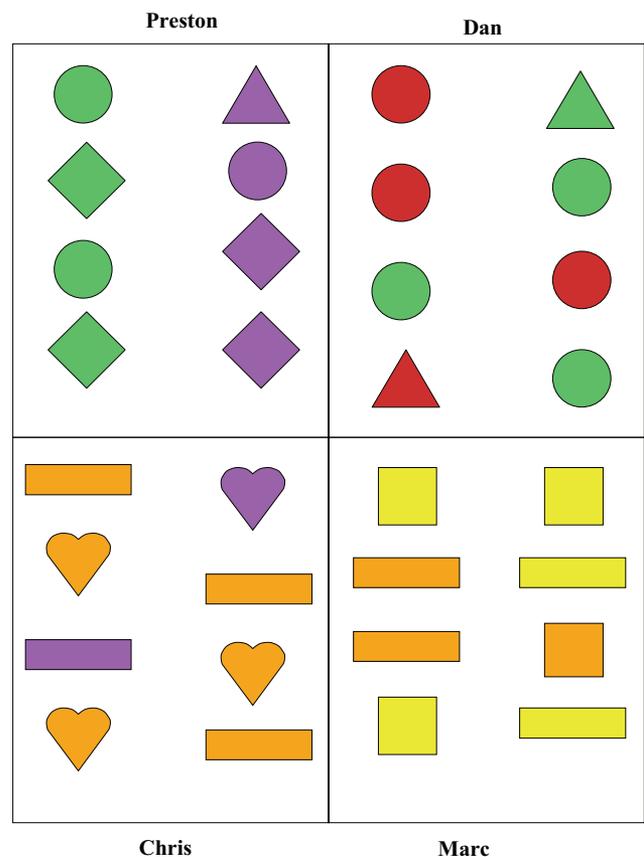


Figure 1. A sample display from the 50–50 structure of Experiment 1. The critical questions would be to predict the color of a new circle drawn by one of these children or to predict the shape of an orange figure. In this condition, the answer derived from the most likely target category is ambiguous.

objects (e.g., different types of animals) or their maker (children who made the drawings). Subjects are told that these displays contain representative samples of the categories shown.

In this paradigm, a new item is partially described. This description typically matches one category well but also matches a second category to a lesser degree. Subjects are then asked to infer some of the object's missing information. For example, they might be told the shape of a geometric figure and asked to predict what color or shading it has. Parallel to the real estate agent example, the secondary category is constructed so that it should either raise or lower the probability of a given feature. For example, the target category could suggest that the geometric figure is likely to be green. In different conditions, the alternative category is either predominantly green or red. Therefore, if people attend to this alternative category at all in making their predictions, they should raise or lower their probability of predicting green, in line with the alternative category. Again, we fail to find this effect in most situations (Murphy & Ross, 1994, 2005; Verde, Murphy, & Ross, 2005). Most of the time, people seem to identify the most likely category and then make their prediction based on that category without integrating information from the alternative category.

Evaluation of Past Paradigms

This past work has provided interesting data on people's category-based inductions. However, it has a number of limitations. One is that evidence for the single-category strategy is based on finding no effect of the secondary category. We have addressed this in a number of ways, such as independently assessing predictions for each category, calculating what the Bayesian model predicts, and then finding that our experiments had the necessary power to find such an effect (e.g., Ross & Murphy, 1996). In other experiments, we have compared the effect of variations of the secondary category with those of the primary category. We found that people are very sensitive to changes in the base rate of features in the target category but not sensitive to an equivalent change in the alternative category (Murphy & Ross, 1994, Experiment 4; 2005), consistent with the claim that people focus on a single category when making inductions. Nonetheless, it is awkward that the evidence for the single-category strategy relies predominantly on a null finding, the lack of any effect of variations in the alternative category. Null results can arise for a number of reasons. Furthermore, this technique results in a binary decision: Either alternative categories are used (there is a significant effect of alternative categories) or they are not (there is no such effect). If they are used more and less in different conditions, there is no simple way to measure the size of the effect in this method.

A second concern about the previous experiments is that only group results are interpretable, making it impossible to identify individuals who might differ from the majority. Consider, for example, the real estate scenario. Some subjects read this scenario with a burglar as the alternative category, and others read it with the cable TV worker as the alternative. Multiple-category responding would be shown as a significant difference between these two groups. Our failure to find such differences is evidence against this proposal. However, group differences may not perfectly reflect individual induction strategies. Perhaps a small minority of subjects did consider multiple categories, but their effects were

washed out by a majority of subjects who did not (or who even used an opposing strategy of some kind, i.e., a contrast effect).

This issue is important for two reasons. First, there has been a growing realization within research on reasoning that people differ systematically in how they solve different problems (Stanovich, 1999) and that individuals differ in their susceptibility to different kinds of cognitive illusions and their ability to resist tempting but incorrect answers (e.g., Frederick, 2005; Stanovich, 2009). In our task, focusing on a single category is obviously simpler and more tempting than using multiple categories.

Second, there is a difference between stating that induction does or does not involve multiple categories and stating that one subset of the population uses multiple categories and another does not. Proposals that visual perception is Bayesian (i.e., decisions are integrated across the predictions of different cues with different reliabilities) rely on the computational requirements of perception and very general assumptions about how visual information is detected and integrated (Kersten et al., 2004). It would be surprising if people differed greatly in such basic processing principles. In contrast, higher-level reasoning processes are more likely to vary across the population, and individuals can solve a given class of problems differently across trials. Indeed, developing competence in a cognitive skill is often characterized not by a sudden shift from one strategy to a more advanced one but rather by a gradual change in the distribution of which strategies are used (Siegler, 1995; Siegler & Crowley, 1991). Thus, variability is to be expected across and within individuals in higher level cognition.

Claims of multiple-category strategies, then, could take either form: One could be proposing that the process underlying induction generally involves multiple categories or one could be proposing that this is one strategy by which people may attempt to perform category-based induction, among others. Our results suggest that the first claim is not tenable. However, the second version of the claim has not been adequately tested. Could Anderson's (1991) model correctly characterize some subset of inductions?

The Present Research

The current experiments advance previous work in a number of ways. First, we present a novel paradigm in which individual responses can be classified as using only the target category or else multiple categories.¹ Therefore, each subject, and indeed (in most conditions) each trial, can be identified with one strategy or the other. As a result, failure to use multiple categories is no longer measured by a null result. Second, we can now compare the amount of multiple-category responding across conditions, which was impossible in the previous, binary method. Third, this new method allows us to classify individuals to determine whether there are consistent differences in how people perform such inductions. Our interest in individual differences is not to try to explain why some people do and others do not use multiple categories. It arises because the existence of strong individual differences will indicate that integrating across categories is not part of the underlying process of induction but rather is a strategy

¹ Experiment 1 was carried out in 2005. Since then, we began a collaboration with Brett Hayes, and he has used the paradigm to investigate other interesting issues, now reported in Hayes and Newell (2009).

that is used depending on task and person variables. This gives a very different picture of the induction process than the one given by previous Bayesian models (Anderson, 1991; Tenenbaum & Griffiths, 2001).

Experiment 1

This experiment used a forced-choice methodology to discover the conditions under which people use multiple categories in predicting features. Three different category structures of geometric shapes ("children's drawings") were used, as shown in Figures 1, 2, and 3. The primary dependent measure was the value of the predicted feature (e.g., what shape or color the item had). As described in detail below, the choice of features would be different depending on whether people relied on a single category (child) or multiple categories. Unlike previous research that required comparing the probability judgments of different groups, in this paradigm, the selection of one feature instead of another gave direct evidence for which strategy a subject used. Therefore, this experiment tests the possibility that past findings of single-category use were a result of group averaging and also discovers whether there are significant individual differences in induction strategies.

Every display was the basis for two critical predictions: one prediction of shape based on color and one prediction of color based on shape. By asking four questions about each of three category structures, we were able to assess for each subject whether he or she was following a consistent strategy across examples for each individual structure type, unlike past research on this topic. We compared three conditions, which differed in how much they encouraged multiple-category responding. In one condition, the target category did not give a clear answer to the induction question. We refer to this as the 50–50 condition, reflecting the equal prevalence of two predictive features in the target category. We suspected that this ambiguity would encourage or even force people to consider the alternative category to break the tie. In another condition (the 80–20 condition), the target category had a clear predominance of one feature, so that multiple categories would not be necessary to make an induction. The third category (the 60–40 condition) was mildly encouraging of multiple-category use, as explained below.

Experiment 1, then, focuses on introducing this new method, measuring whether some individuals use multiple categories in induction, and examining how the prevalence of this strategy may change with ambiguity, both in general and individually. Experiment 2 uses this method to test an explanation of these results.

General Method

Participants. The subjects were 48 New York University undergraduates who received course credit for participating. One subject was removed from the study because of excessive classification errors.

Materials. The items were described as drawings made by children using a simple drawing program. Each page contained the drawings of four children, presented in four labeled quadrants. The drawings used easily recognized shapes and colors. Details of category construction are described separately for each condition below. The pages were printed in color, placed in plastic sleeves,

and collected into six-page booklets. Subjects answered the induction questions in separate pamphlets, where each page of the pamphlet corresponded to one booklet page.

There were four versions of the booklets/pamphlets, differing in order of questions (see Procedure section of this experiment) and feature counterbalancing. For example, if one version of the category had hearts as the multiple-category response and diamonds as the alternative response, the other version reversed these two. Therefore, across subjects, any particular strategy that might be used was not confounded with preference for particular shapes or colors.

The pamphlets consisted of six pages, each containing three sets of questions about a given category set. Each set contained three questions in the following pattern:

I have a purple figure. Which child do you think drew it?

What is the probability (0–100) that the child you just named drew this?

What shape do you think the figure has? (Circle one)

Underneath the final question was an alphabetical list of colors or shapes present in the display. This choice was the primary dependent measure. Overall, there were 18 trials: two critical sets of questions and one filler set for each display. The fillers were intended to obscure the abstract similarity of the critical questions.

Procedure. Subjects read instructions explaining the cover story in which (imaginary) children had made drawings differing in color and shape. The displays were said to contain a representative sample of the drawings. The instructions explained that "each child had definite preferences for the shapes and colors they like to draw." Subjects were asked to look at the first page of the booklet to try to learn what each child's drawings were like and how they could be told apart. After doing so, they received a second page of instructions for the induction task. The second set of instructions explained that they would be told about part of a drawing and were then to decide who they thought drew it and what the rest of the drawing looked like. They were also instructed in the use of the 0–100% probability scale, in which "0% means that something is impossible (would never happen), and 100% means that it is completely certain (would always happen in this situation). 50% means that the thing would happen about half the time."

Subjects answered the questions about the first page of the booklet and then went through the remaining pages and questions at their own pace. The categories were readily available for viewing the entire time they were answering the questions. Half of the subjects completed problems in the order indicated below (50–50, 60–40, 80–20), and half followed the reverse order. Noticeable order effects occurred in only one condition, which we describe below.

Category Structures and Results

Because the category structures are complex and require detailed explanation, we describe each one in turn, along with its results, rather than presenting the Method and Results separately as usual.

The 50–50 structure: Method. In the first structure, shown in Figure 1, there was an even split between the predicted features within the target category, so that a definitive answer could not be given without using multiple categories. For example, consider the question of predicting the color of a new circle. Because Dan has drawn more circles than anyone, he represents the target category. However, half of Dan's drawings are red and half are green (and the same is true of Dan's circles). Because an unambiguous answer cannot be given from a single category (Dan), this situation might encourage subjects to consider the alternative category, represented by Preston, who has also drawn some circles. If subjects were to consider this alternative category, they would find additional evidence for green drawings but none for red drawings. Thus, they would be more likely to choose green as the answer. The other question was to predict the shape of an orange figure, where half of Chris's figures are hearts and half are rectangles. Considering Marc's drawings would give further evidence for rectangles. Thus, predicting the color of a new circle or the shape of a new orange figure are the two critical problems for this display.

According to Anderson's multiple-category rule, people go through the following steps in predicting the color of a circle. First, they estimate the likelihood that this circle is in each category: $P(\text{Dan}|\text{circle}) = 6/9 = .67$; $P(\text{Preston}|\text{circle}) = 3/9 = .33$. The other categories have probabilities of 0, so can be ignored for the remaining computations. Next, subjects must calculate the probability of different predicted features (colors), given each category:

$$P(\text{green}|\text{Dan}) = 4/8 = .5;$$

$$P(\text{green}|\text{Preston}) = 4/8 = .5;$$

$$P(\text{red}|\text{Dan}) = .5;$$

$$P(\text{red}|\text{Preston}) = 0.$$

According to this rule, then,

$$\begin{aligned} P(\text{green}) &= P(\text{Dan}|\text{circle}) \times P(\text{green}|\text{Dan}) + P(\text{Preston}|\text{circle}) \\ &\quad \times P(\text{green}|\text{Preston}) \\ &= .67 \times .5 + .33 \times .5 = .5 \text{ and} \end{aligned}$$

$$\begin{aligned} P(\text{red}) &= P(\text{Dan}|\text{circle}) \times P(\text{red}|\text{Dan}) + P(\text{Preston}|\text{circle}) \\ &\quad \times P(\text{red}|\text{Preston}) \\ &= .67 \times .5 + .33 \times 0 = .33. \end{aligned}$$

Therefore, if they use multiple categories, subjects should be more likely to respond with the color green than with the color red. These relations held for all of the inferences in the 50–50 category structure. Although these probability computations may seem rather difficult for people to do, they encompass a number of different strategies, such as looking at Dan, being uncertain, and then looking at Preston and seeing additional green circles. Subjects do not need to do these actual computations in order to predict the color green—they simply need to consult both categories. The main point is that using multiple categories will lead to a consistent answer of green. (These predictions assume a deterministic response rule, in which if green is more probable

than red, people will therefore choose green as their answer. We consider other response rules in the Response Distributions section below.)

The 50–50 structure: Results. Subjects' mean confidence in their initial category judgment was 67.1%, indicating that they were not certain that the item was in the target category. (In fact, there were no ratings of 100% certainty in this experiment.) This figure is also close to the 6/9 figure of the probability computation above. Given that subjects were not certain of the item's category, on normative principles they should have paid attention to alternative categories in making predictions. In fact, they did choose the feature indicated by the alternative category 59% of the time, which was marginally different from chance, $t(46) = 1.82$, $p < .10$. An analysis of individual subjects examined whether they consistently chose the feature indicated by the alternative category (4 of 4 times), chose the opposite (0 of 4 times), or fell in between. (Although choosing the opposite consistently would be perverse in this structure, it will not be in future structures.) Giving the same answer all four times would be expected to happen by chance only .0625 of the time for either type of answer (very conservatively assuming that subjects are selecting among the two critical features only, even though there are more present.). This analysis showed that 13 of 47 subjects consistently chose the feature indicated by the alternative category, which is significantly more often than would be expected by chance according to a z -test approximation of the binomial distribution (with Yates's correction), $z = 5.76$, $p < .001$. Only three subjects consistently chose the opposite of the expected feature, which is what would be expected by chance ($.0625 \times 47 = 2.9$).²

This pattern of results suggests that when the prediction was ambiguous in the target category, a subset of subjects (28%) consistently used alternative categories to determine the answer. In this structure, then, it seems that a significant minority of the population follows something like Anderson's rule. The overall selection of the multiple-category feature (e.g., green, for the circle example considered) was not very high in an absolute sense—59%, when use of a single category would lead to 50% selection of that feature—but this is still definite multiple-category responding that we had not found in earlier studies with such materials. Notably, the analyses show that this overall figure results from a mixture of subjects who consistently use multiple categories with a larger number who do not.

The 60–40 structure. The next design had less ambiguity in the target category prediction. In Figure 2, subjects were asked to imagine that a diamond had been found and to predict its color. Here, Drew is the target category, and Julian is the secondary category. This design is similar to the 50–50 structure, except that there is now a slight predominance of one color in the target

² The predictions stated above consider two possible answers each subject might give: either the one predicted by Anderson's model or the one using the target category alone. However, subjects could, and sometimes did, give other answers. In almost every case, this answer was a different feature that was present (once) in the target category. (In fact there are no such features in the 50–50 structure, but there were in other structures.) Therefore, these were coded as single-category responses. One other answer was seemingly arbitrary, or perhaps an error in circling the correct feature, as it was not present in the target category at all; this response was deleted from the analysis.

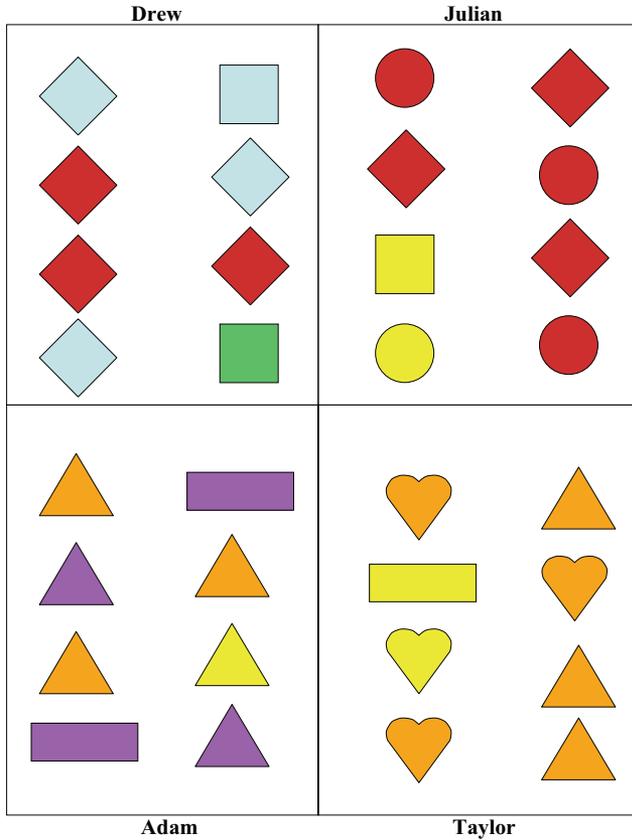


Figure 2. A sample display from the 60–40 structure of Experiment 1. The critical questions would be to predict the color of a new diamond drawn by one of these children or to predict the shape of an orange figure.

category. That is, Drew has four blue figures and three red figures (hence, it should be called the 57–43 design, but we rounded). If one looks only at diamonds (the given feature), however, the split is still 50–50: three reds and three blues. Thus, if people are only considering items that have the given feature (as some people undoubtedly do; Murphy & Ross, in press; Papadopoulos, Hayes, & Newell, 2009), then they will not be able to make a clear decision and will likely examine Julian’s diamonds (as found in the 50–50 design). In contrast, if subjects are considering the entire target category (thinking, “Drew made more blue items than anything else”), then there is no ambiguity, and they will choose the predominant feature in the target item. Therefore, this design encourages multiple-category responding to some degree (i.e., for people who focus on the given feature), but not as much as the previous structure does, where the target category provided no way to choose between the features.

Anderson’s model makes the following predictions:

$$\begin{aligned}
 P(\text{blue}|\text{diamond}) &= P(\text{Drew}|\text{diamond}) \times P(\text{blue}|\text{Drew}) \\
 &\quad + P(\text{Julian}|\text{diamond}) \times P(\text{blue}|\text{Julian}) \\
 &= .67 \times .5 + .33 \times 0 = .33.
 \end{aligned}$$

And correspondingly for red, we obtain:

$$P(\text{red}|\text{diamond}) = .67 \times .375 + .33 \times .75 = .50.$$

Thus, although blue wins when one looks within the target category (.5 vs. .375), red wins when integrating across categories.

The 60–40 structure: Results. As in the previous structure, participants did not express strong confidence in their initial categorization choice (67.3%), suggesting that they should have considered the alternative category. However, there was little overall tendency for subjects to select the property consistent with multiple categories, which they did only 53% of the time, a level of performance that was not different from a chance level of 50%, $t(46) = .45, p > .50$. Notably, this middling result disguises strong individual patterns, as there were more subjects than expected from chance using both strategies: A total of 16 of 47 subjects used multiple-category induction on 4 of 4 trials, $z = 7.57, p < .001$, and 12 of 47 consistently chose the predominant feature in the target category (single-category induction), $z = 5.16, p < .001$.

The first result is similar to that of the previous structure; what is new is the second result, in which a statistically unlikely number of people ignored the alternative category and chose the predominant feature in the target category. That is, the small reduction in ambiguity in the primary category led to a stronger reliance on that category. In fact, the increase in single-category responding from the 50–50 to the 60–40 design (from 3 to 12 subjects) was significant by a Fisher’s Exact Test, $p < .05$. This increase shows that at least some subjects were paying close attention to the category structure. If they had been looking only at diamonds, for example, without regard to the category in which they were located, they would have chosen the alternative feature (red). The subjects who consistently chose the target category feature on all questions were obviously focusing on a specific category. The subjects who consistently chose the alternative feature must have been using multiple categories.

The 80–20 structure: Method. The final condition greatly reduces the ambiguity of the target category prediction: When people consult the target category, they will arrive at a clear answer. Consider Figure 3 and imagine that you have been presented with a new square. Most likely Federico drew it. Notice that four of Federico’s figures are green and three are purple; and the same is true for the squares. The latter is unlike the previous structure, in which the prediction was ambiguous in the items with the given feature (squares, in this case). Now one might expect even more people to pay attention to only Federico’s drawings (the single-category strategy). However, if they used information about Cyrus, who has four squares and many purple drawings, they would instead choose purple as the most likely feature (the multiple-category response), regardless of whether they confined themselves to squares.

In Anderson’s model, the strong evidence for purple in Cyrus’s drawings is enough to outweigh the slight advantage for green in the target category:

$$\begin{aligned}
 P(\text{green}|\text{square}) &= P(\text{Federico}|\text{square}) \times P(\text{green}|\text{Federico}) \\
 &\quad + P(\text{Cyrus}|\text{square}) \times P(\text{green}|\text{Cyrus}) \\
 &= .64 \times .5 + .36 \times 0 = .32 \text{ and}
 \end{aligned}$$

$$P(\text{purple}|\text{square}) = .64 \times .375 + .36 \times .75 = .51$$

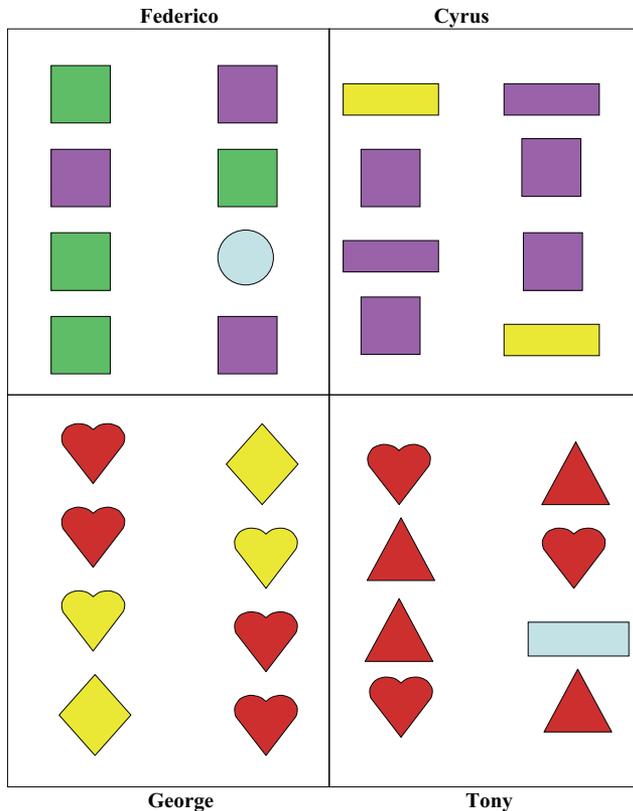


Figure 3. A sample display from the 80–20 structure of Experiment 1. The critical questions would be to predict the color of a new square drawn by one of these children or to predict the shape of a red figure. In this condition, the prediction is unambiguous.

Thus, using both categories would yield an advantage for purple. These numbers are virtually identical to the predictions for the 60–40 design, so Anderson’s model predicts similar results.

The 80–20 structure: Results. In this condition, people estimated 64.7% confidence in their initial categorization, so they should have used the alternative categories. In this condition, where the target category gave an unambiguous induction, the percentage of subjects choosing the multiple-category feature (purple) declined to 30%, which is significantly below chance, $t(47) = -3.7, p < .01$. The individual-subjects analysis of this design shows the reverse pattern from the 50–50 condition: Only seven subjects participants consistently followed Anderson’s rule, choosing purple (or its equivalent) all the time $z = 2.14, p = .08$, and 22 consistently chose the feature from the target category (the single-category response), far more than would have been expected by chance $z = 11.18, p < .001$. The latter result was the strongest pattern in any of the forms tested. Thus, it seems that most people were unlikely to use the alternative category when there was a clear answer within the target category. That is, the majority of responses were based on the target category, with 47% of the participants consistently following a single-category strategy. Nonetheless, about 15% of participants consistently used multiple categories even in this condition.

One uninteresting explanation of the difference between those who used a single category and those who used multiple categories would be that the former were very confident in their classifications and therefore considered only one category. If a subject thought that the figure was 95% likely to be drawn by Federico, then he or she would be perfectly justified in not letting other children’s drawings influence the induction. However, the subjects who consistently used a single category had the same confidence in their classifications as those who consistently used two categories ($M_s = 63.9%$ for both groups). Clearly, those who used a single category did not do so because they were certain that this category was correct.

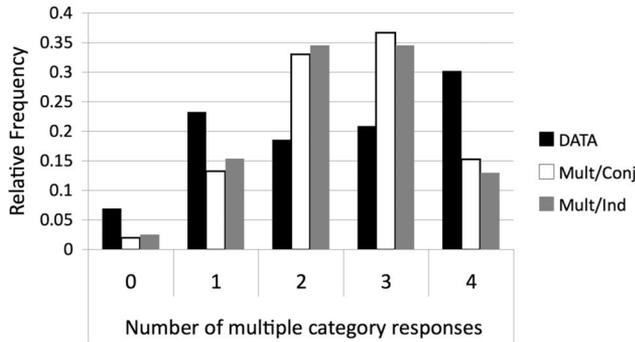
Response Distributions

Our data analysis has focused primarily on the mean level of multiple-category choice and people who seem to be following a consistent strategy. However, an alternative explanation of our results requires us to consider the detailed distributions of responses. We have been assuming a deterministic response rule in which people choose the feature with the highest likelihood. However, it is possible that, instead, people probability match. Imagine that people think that the probability that a target feature is green is .50 and the probability that it is red is .30 (less likely possibilities being ignored). Rather than simply choosing green as we assumed, they might choose green $.50 / (.50 + .30) = .625$ of the time and red $.30 / (.50 + .30) = .375$ of the time. But choosing an answer with .625 probability across four trials could result in that response from zero to four times. Thus, to evaluate those rules fully, we need to consider the distribution of responses across subjects. In all of these calculations, we consider only the two most likely features, because people very seldom selected any other features, which had much lower probabilities. Those choices probably reflect uncooperative responses.

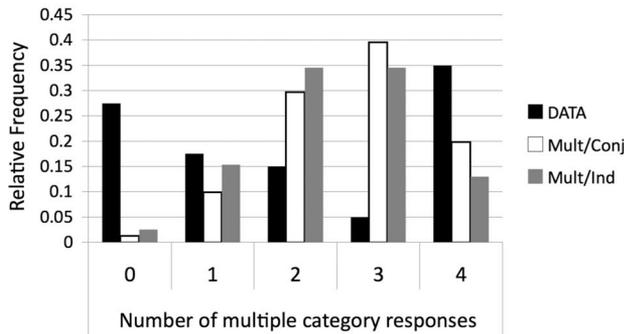
We calculated predictions for Anderson’s (1991) model, assuming probability matching, for each of the three category structures. Using the binomial theorem, one can derive predictions for how often people should make 0, 1, 2, 3, or 4 multiple-category responses given a certain probability. For example, using the figures above, the chance of obtaining a response like red 4 out of 4 times would be $.375^4$ or .02. More likely would be a distribution of 3 greens and 1 red = .37. We derived such predictions for two multiple-category theories. The first assumed independence, that is, Anderson’s (1991) model. The second assumed that people only examine items that have the given feature (see Murphy & Ross, in press; Papadopoulos et al., 2009), that is, that they are looking for conjunctions of features. We call these the *multiple/independent* and *multiple/conjunction* models, respectively. As a general rule, these models predict that not many people will follow a consistent strategy (4 or 0 trials of the same kind of response). The Appendix derives the predictions for each of these multiple-category models (as well as the single-category models) for each condition.

Figure 4 shows the distributions that are predicted by these two multiple-category models for each of our three category structures, along with the observed distributions (including only those subjects who provided four data points). In the 50–50 structure, far

A: 50/50 Condition



B: 60/40 Condition



C: 80/20 Condition

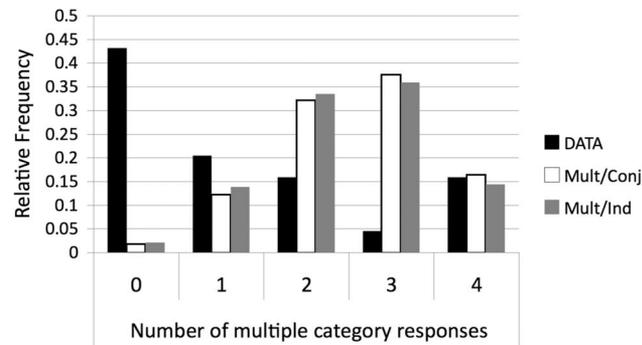


Figure 4. Distributions of responses by subjects and predictions of two multiple-category models assuming probability matching: one assuming feature independence (Mult/Ind) and one assuming that only items having the given feature are used, that is, conjunction seeking (Mult/Conj). The data include all subjects in each condition who classified all four of the trials correctly ($n_s = 43, 40,$ and 44 for conditions 50/50, 60/40, and 80/20, respectively).

more subjects consistently chose the multiple-category response than the probability-matching models predict (rightmost bars). In the 60–40 condition, there are far more people in *both* endpoints than the models predict. That is, some people consistently avoid using multiple categories and others always use them. Finally, in the 80–20 condition, a very large number of subjects never use multiple categories—far more than can be predicted by probability

matching. Chi-square tests showed highly significant differences between the observed distributions and those predicted by probability matching, but because of low expected frequencies in some cells, these tests are not fully appropriate. However, the main point is not the quantitative fit of the probability-matching models, but the fact that they cannot qualitatively predict the consistent responses we observed.

In short, although our use of deterministic response rules was a simplification, it does, in fact, provide a better account of the data than does a probability-matching rule. Most notably, the 60–40 distribution (Figure 4) strongly suggests a mixture of different subjects who were consistently following one or the other strategy, resulting in a bimodal distribution that is essentially the inverse of the probability-matching predictions. Rather than one model with probability matching being able to account for these results, it seems likely that we need to posit two distinct strategies—single-category and multiple-category—that are differentially promoted by the different category structures. Even more sophisticated unitary models that might fall between the extremes of probability matching and completely deterministic responding cannot predict a high proportion of subjects consistently following different strategies.

Discussion

The results of this experiment show that people use multiple categories under some circumstances, but when the most likely category of an uncertain object gives a clear answer (the 80–20 structure), they generally do not do so. Although subjects’ ratings showed that they were not certain of which category the object was in, if the most likely category had a predominant color or shape, subjects generally attributed that color or shape to the object, rather than considering other categories. However, when the target category was somewhat (60–40) or completely (50–50) ambiguous about the induction feature, people were increasingly likely to use multiple categories.

Using the medical analogy from the introduction, this would mean that if your doctor knew an obvious treatment for your suspected infection (i.e., predicted that your condition would improve from this treatment), she might not consider whether the treatment would work for the less likely alternative, the virus. However, if there were two equally effective treatments for the infection, she might think, “Treatment X would also help the virus, so why don’t I prescribe that?” Of course, one might expect medical professionals to consider multiple possible diagnoses under all conditions, unlike our naive subjects. We consider this issue in the General Discussion.

It is striking that even in the most ambiguous case, people did not overwhelmingly use multiple categories. In the 50–50 condition, there were two equally frequent colors, say, red and green. If people looked at the other possible category, they would find strong evidence for red, say. However, only 59% of the responses were red—slightly more than chance—and only 28% of subjects consistently used this strategy. Therefore, although the findings for this condition provided evidence that people can and do use multiple categories, it seems that most subjects failed to do so.

The 80–20 category structure, which was unambiguous in the target category, yielded even fewer multiple-category responses. Overall, only 30% of responses consisted of the multiple-category

feature,³ and only 7 (15%) of 47 subjects consistently chose this response. This condition was the only one that had a noticeable effect of question order. When it was presented first, only 23% of responses were multiple category, whereas when it was last, 38% were. (Five of the seven consistent responders came from the latter order.) Although these order groups were not significantly different, $t(45) = 1.29$, $p > .15$, this trend suggests that at least some of the multiple-category responders in this condition got the idea from the earlier conditions that especially encouraged it. That is, the 30% overall multiple-category responding could be something of an overestimate of the effect in naive subjects.

In short, these results support our previous conclusions (Murphy & Ross, 1994; Ross & Murphy, 1996) that people have a strong tendency to focus on a single category when making inductions in this task. Importantly, this evidence is no longer in the form of failure to find an effect of alternative categories but is rather a positive finding of people choosing the target-category feature rather than the feature predicted by a multiple-category use.

A new facet of the results is the marked individual differences, which could not be discovered in our past designs. All three conditions contained more consistent responders than would be expected by chance. The 60–40 condition was perhaps the most interesting one, because it contained significant numbers of both single-category and multiple-category responders. The strong differences in strategy use were driven by both the category structures and individual differences in reasoning strategies. In fact, there were five subjects who gave all multiple-category responses in all three conditions (12 critical trials). It seems most likely that these subjects consciously worked out something like Anderson's rule for making inductions and consistently followed it. However, most subjects either were not consistently strategic or varied which strategy they used across conditions.

Finally, the results speak to an alternative explanation of results from this sort of paradigm (which has been suggested to us in talks and other informal settings). The alternative suggests that people do not use the categories but merely count up the conjunctions of the given and target features. For example, in Figure 3, when asked to predict the color of a new square, people would simply count that most are purple, ignoring who drew them. However, this explanation cannot account for many of the important effects. First, as we noted in the 80–20 condition shown in Figure 3, this strategy leads to an answer of purple (the multiple-category response), because most squares are colored purple. However, the dominant response in this condition was green—the color most common within the target category. Furthermore, the different patterns of responses found across the three conditions can only be explained if people attend to the target category as being of particular importance. People often give the multiple-category response in the 50–50 condition because the target category is ambiguous, but they give the multiple-category response very seldom in the 80–20 condition, because the target category gives a clear answer. If the subjects had ignored the categories, they would have given the same answer in all conditions. Thus, most subjects in the present experiments are clearly using the categories.⁴

It may be that our practice of asking a categorization question prior to the induction question increases the use of the target category (Hayes & Newell, 2009). Although the categorization question may cause people to focus more on the category they

choose, this is not a very rational response, because subjects also indicate in their ratings that they are not certain about the category. If you have just written that Federico is the most likely category, but you give it a probability of only 66%, then you presumably are aware that Cyrus is also a potential category (or else you would have rated Federico's likelihood as near 100%). The probability rating requires subjects to acknowledge their uncertainty. Therefore, it is even more striking that they answer the very next question as if they had just written 100% rather than 66%. We explored this puzzle in the next experiment.

One reason that we use the initial categorization question is because the categories are unfamiliar and perhaps not entirely convincing ones. In real life, people certainly believe that cars are different from sport utility vehicles and that cats are different from squirrels. The categories in the present experiment are clearly fictional. We ask the classification question in part to keep the categories in participants' minds. In everyday life categories, one classifies objects without any special instructions or effort. Indeed, it is probably very difficult to ignore familiar categories if one tries to do so.

Experiment 2

Previously, we have explained people's tendency to use a single category in induction by the proposal that people answer the induction problem in two separate stages (Murphy & Ross, 2007): First, they decide which category the item is in, and then they make the induction based on properties of that category. The problem is that people generally treat these two stages separately, such that any uncertainty in the first stage does not carry over to influence the induction process of the second stage, leading to suboptimal responses. People may think something like, "A red figure . . . it could be drawn by Tony or George. Tony has more red ones, so it's more likely to be him, I guess. [Circles "Tony" on form and writes "67%.]" So, what shape is it? Well, Tony's figures are mostly triangles . . ." That is, although multiple categories are considered in the first stage (as shown by people's probability judgments on the classification), when making the induction, only the most likely category is considered. Why is this?

From the perspective of the psychology of reasoning, people's use of a single category is not as surprising as it might first seem. In an extensive review, Evans (2007) argued that one of the central principles of higher level reasoning is *the singularity principle*: When reasoning hypothetically, people consider only one situation at a time. They consider alternative situations only if the first one does not lead to an answer or if an external factor such as a hint causes them to do so. A related notion is found in mental model theory, which claims that errors in syllogistic reasoning often result from people considering one mental model and not seeking alternative ones that are also consistent with the premises

³ Note that this figure cannot be directly compared with the proportion in the 50–50 structure, in which half of the people using a single category could have given the multiple-category answer simply by chance. Here, there is no need to guess, as the target category gives a clear induction.

⁴ In the present category structures, ignoring the categories yields the same inductions as the multiple/conjunction strategy outlined earlier and shown to be incorrect in Figure 4. However, psychologically these are very different processes, as one of them involves ignoring the categories and the other involves weighting them.

(Johnson-Laird, 2006, Chapter 16). Evans related this principle to a number of different phenomena in logical and statistical reasoning. One that is relevant to the present endeavor is his analysis of whether people attend to base rates in Kahneman and Tversky-like problems (Kahneman & Tversky, 1973) in which they are given both base-rate information and diagnostic information about the current case. Although there is controversy over when people do use base rates (e.g., Gigerenzer & Hoffrage, 1995; Koehler, 1996), it seems fair to say that their use of this information does not correspond closely to Bayesian predictions.

As Evans (2007) pointed out, people do use base rates under some conditions, most notably when they are the only information provided (as shown by Kahneman & Tversky, 1973, p. 242). Evans (2007) argued that the problem people have is in integrating base-rate information with information about the specific object being evaluated. "Instead, they base their mental simulations *either* on the base rate or (more frequently) on the diagnostic data in line with the singularity principle" (p. 143; see also Evans, Handley, Over, & Perham, 2002). Related to this principle is Stanovich's (2009) characterization of people as *cognitive misers*, that is, as trying to get away with the least amount of explicit reasoning possible. If people can answer a question based on diagnostic information, they see no need to consider other sources of information such as base rates.

This analysis is similar to our own explanation of single-category reasoning in the category-based induction task. First, both accounts suggest that people understand abstractly that two sources of information are relevant, but they nonetheless do not use both. That is, although subjects in our task know that multiple categories are potentially relevant, once one category has been selected, it does not occur to them to look at a second one. Second, we both claim that this focus on a single piece of information is a default option that can be overridden in the right circumstances. For example, in Experiment 1, when one category gave an ambiguous prediction, at least some subjects then attended to another category to make the prediction, rather than choosing randomly. This corresponds to Evans (2007) proposal that singularity can be overridden when one situation does not lead to an answer. Hints and reminders to consider other possibilities can cause people to choose better answers in reasoning tasks (Stanovich, 2009).

We explored this explanation of single-category reasoning in a second experiment by changing the initial classification question so as to discourage people from focusing on a single category. Experiment 2 followed the general procedure of the first experiment, again using categories of children's drawings. However, rather than asking subjects to guess the most likely category of the item and judging its probability prior to the induction, we listed all four categories and asked subjects to write down the probability that the item was in each category, requiring these numbers to sum to 100%. This procedure was expected to have two important differences from the earlier experiment. First, it does not involve selecting any category as the most likely one and thereby does not create as obvious a candidate to be the single category considered in the induction stage. Second, by making nonzero responses to one of the other categories, this category should be brought to mind and therefore is more likely to be used in induction.

This experiment is almost identical to Experiment 1 in the information subjects had to consider. If a subject were told about a red drawing and chose Tony as the category with probability

60%, this implies that other categories must have 40% of the likelihood. Since only one other category contained red figures, subjects could have easily considered both categories. In the present experiment, that subject would presumably write down 60% for Tony and 40% for George. Thus, although in both cases subjects must implicitly consider both categories in order to answer the questions, in Experiment 2 they must also explicitly acknowledge the relevance of George and will not have explicitly chosen Tony as "most likely." These changes may reduce the reliance on a single category.

Method

Thirty-eight New York University students served as subjects for course credit. The experiment was identical in design, categories, and procedure to that of Experiment 1. However, the initial questions concerning the test item no longer required a single classification. Instead, the question took the following form:

I have a purple figure. What is the probability that each child drew it?

Preston ___% Dan ___% Chris ___% Marc ___% (Must sum to 100.)

Following this, the induction question asked subjects to predict the unknown feature (here, shape) by circling one value, as in Experiment 1.

Results

Two subjects gave repeated answers of features that were very uncommon in the categories; the subjects were dropped from the experiment as being uncooperative or confused. Recall that in Experiment 1, we eliminated trials in which subjects did not choose the intended target category as the most likely one. In the present experiment, people did not circle the target category, so we included any trial in which the probability written down for the target category was the highest or tied for most likely. The tied cases may indicate a strategy in which subjects are considering multiple categories, only not weighing them by their correct probabilities. Most of the critical trials had two categories with nonzero responses, as intended, and 27 of 38 subjects always chose the two intended categories as nonzero responses.⁵ A few subjects seemed suspicious of zero probabilities and so gave nonzero probabilities to all four categories (which may technically be correct, as the probability is surely not truly 0, even when there is no evidence that a child had drawn a particular feature in the past). On average, subjects selected 2.12 categories as being possible categorizations for the test item, so their inductions should have used multiple categories.

The main question of interest is how often subjects made multiple-category property inductions with this new methodology. In striking contrast to Experiment 1, 89% of the inductions were

⁵ During analysis, we discovered an error in one display, which was intended to have seven triangles in the target category and four in the secondary category. An additional triangle was placed by mistake into a third category. Therefore, some subjects gave nonzero (low) probabilities to this category. We omit this item from the calculations of how many categories subjects gave nonzero responses to, as the correct answer was not two. The error did not influence the induction responses to this question, which were identical to the overall pattern of results.

multiple-category responses across the three conditions. Indeed, 22 of the 38 subjects *always* gave the multiple-category response. (Recall that in Experiment 1, with the same categories and inductions, only 5 of 47 subjects always used multiple categories.) This high consistency of responding did not allow for much difference among the three conditions that were so important in Experiment 1, although their means followed the previous order: The mean multiple-category responses were 90%, 89%, and 87% in the 50–50, 60–40, and 80–20 conditions (compared with 59%, 53%, and 30% in Experiment 1). Clearly, the most important result is that asking subjects to list the probabilities for all four categories causes them to then use multiple categories in induction, and the remaining differences among conditions are trivial.

Discussion

The contrast between the present results and those of Experiment 1 could not be more marked. This is best seen in the 80–20 condition, where subjects in Experiment 1 made the multiple-category response only 30% of the time, compared with 87% of the time in the present experiment. Indeed, the single-category responses in Experiment 2 are predominantly attributable to four subjects who stuck to this strategy fairly strongly. When these participants are dropped from the analyses, the overall rate of multiple-category responses is fully 95%. Furthermore, more than half of the present subjects gave the multiple-category response on every single question (of 12), which was more than in any of the three conditions of Experiment 1.

Clearly, asking subjects to provide probabilities for all categories led them to not focus on the most likely category to the exclusion of others—that is, it led them to integrate information across categories. Note that this difference could not have been found with our previous between-subjects method. At most, we would have discovered a significant difference in the present case, and (perhaps) a null result in Experiment 1. Notably, the present method, in which individual responses can be identified with a given strategy, allows us to see that this procedural change results in widespread use of multiple categories, rather than a switch by a few, perhaps more intelligent, subjects. These findings give strong support for the two-stage model we have proposed.

One might wonder why this technique gave such strong evidence of multiple-category use, when a number of our earlier experiments using similar manipulations had little effect. For example, in Malt et al. (1995) and Ross and Murphy (1996), which used scenarios, we asked about the categorization of the critical object only after the induction question had been answered. Therefore, people had not explicitly selected a target category when doing the induction, and yet they showed no sign of multiple-category use. Similarly, in experiments using visual displays like the present one, we have omitted the classification question (Murphy & Ross, 1994, Experiments 2 and 5) or placed it after the induction questions (Murphy & Ross, 2005, Experiment 4)—each time failing to find evidence of multiple-category use. Thus, although the initial question may have some effect (Hayes & Newell, 2009), it does not seem that its mere presence or absence is responsible for multiple-category use. People may often identify a most likely category even if they are not asked to do so.

Instead, we believe that the particular method used here, in which subjects indicated that multiple categories were relevant, without making an overt decision as to which one was most likely, is the key. In order to overcome people's tendency to focus on a single possibility, that is, to follow the singularity principle, one needs to encourage them to explicitly acknowledge each possible outcome without allowing them to choose one of them. If the classification question is simply omitted, people may spontaneously select the most likely category and then ignore the less likely categories. (This seems especially likely when familiar categories are used, as in Ross & Murphy, 1996.) Making subjects indicate that another category is also possible seems to add greatly to the use of multiple categories in induction.

General Discussion

Our past research has shown, using group comparisons, that people generally seem influenced by only a single category when making inductions about an uncertain object. The present experiments help us understand why this is so. Perhaps people only have enough working memory capacity to focus on one category at a time. Or perhaps they simply do not understand the principle of hedging their bets by considering a less likely category. However, such explanations cannot account for the people and situations in which multiple categories are in fact used that we found here.

Our explanation (Murphy & Ross, 2007) is related to Evans's (2007) singularity principle. When people think about an item whose categorization is uncertain, the most likely category is usually the only one considered. This is not because of a positive decision to ignore other categories, but because of a broad tendency to focus on one situation or possibility at a time. So long as that possibility leads to an answer, other possibilities are simply not considered. What is counterintuitive about our result is that people do this even when they have just admitted that they are not certain about their categorization. Equally surprising, a small change in procedure leads people to overcome this tendency. In the original paradigm, if subjects say that a picture is 65% likely to be drawn by Tony, they must have looked at the display, noticed that Tony had the most red drawings but that George had a few as well, and therefore come up with the 65% figure as a relative proportion. The fact that they wrote down a probability less than 100% implies that they identified multiple categories that the stimulus could be in. Yet, when it comes time to make the induction, the 65% likely category is treated as a single uncertain possibility rather than as two possibilities: one 65% likely and the other 35% likely.

In contrast, in Experiment 2, people wrote down the probabilities of two categories, which remained visible just above the induction question. The response did not single out one category. Psychologically, then, there is a surprisingly large difference between thinking about one category that is 65% likely to be correct and one that is 65% likely plus another one that is 35% likely to be correct. In the first case, the singularity principle applies, but in the second, it does not.

In real life, both of these situations may occur in different contexts. In medicine, presenting cases are often consistent with multiple different diagnoses (categories), which doctors attempt to narrow down. However, once a physician begins to focus on an initial diagnosis, the other possibilities may be neglected. Groopman (2007) identified such reasoning as a major source of medical

errors, especially for cases that drag on for months or even years without a satisfactory result. He suggests that when patients are not satisfied with their treatment, they should explicitly ask their doctors, "What else could it be?" or "Is it possible I have more than one problem?" (p. 263). If this technique is indeed effective, it seems related to our findings in Experiment 2. These questions do not provide the doctor with new information, but by placing new possibilities into open consideration, Groopman suggests, patients can help their doctors to take them seriously, much as our subjects considered the secondary category seriously when they wrote "35%" next to it.

Our findings show that the problem in considering multiple options in category-based induction is not that people do not grasp the issue of bet-hedging, in which one adjusts a prediction based on different possible outcomes. They do follow this principle if (a) focusing on a single category does not lead to a clear answer (as in the 50–50 condition in Experiment 1) or (b) they identify multiple categories as being relevant without selecting one of them (as in Experiment 2). Thus, the problem in this kind of task is not that people do not understand that multiple categories are relevant but that default reasoning strategies prevent multiple-category reasoning.

Individual Differences

A new and intriguing result from the present investigation is the finding of large individual differences. Our previous work, which required group comparisons, had no way of assessing whether a minority of subjects were, in fact, using multiple-category reasoning in contrast to the majority who were not. The present results reveal that this was actually the case. Consider the 80–20 condition of Experiment 1, which had the most single-category responding of all our conditions. Even here, seven subjects consistently used two categories, but these were outweighed in the group results by the 22 who consistently gave single-category responses. The mean proportion of multiple-category responding was significantly less than chance, so a group comparison would not have revealed the existence of this minority group of reasoners.

The most striking example of individual differences was the 60–40 condition of Experiment 1, in which the overall rate of multiple-category use was at chance levels, but this hid two distinct patterns in which 34% of subjects always used multiple categories and 25% never did. These strong individual differences suggest that the induction process proposed by Anderson (1991) is not the underlying process of induction but is one of a number of strategies that people may draw on. Which one they use is partly determined by the details of the task—the procedure, materials, and instructions—which, in turn, suggests that most people are able to use different strategies, and they shift based on the task (e.g., compare Experiments 1 and 2). However, the individual differences within each condition suggest that choice of strategy may also reflect cognitive styles or patterns of reasoning that vary across the population. (See Stanovich, 2009, for a detailed discussion of related individual differences in reasoning.)

It is tempting to ascribe the consistent use of multiple categories to the more intelligent or perhaps more educated of the subjects, similar to findings in decision-making research (Stanovich, 1999). Within our subject group, it is possible that those who have taken and mastered statistics and probability would have a better understanding of how to hedge their bets by considering multiple uncertain categories. Finally, since bet-hedging requires more com-

putation, motivational factors might influence who chooses to consider multiple categories. Because we do not have personal information about individual subjects, we cannot test these possibilities, which represent an interesting topic for future research.

In addition, uncertainty can be unpleasant. Sometimes, when waiting for an uncertain outcome—a medical diagnosis, news about a job application, or test scores—one can feel that it would be better to receive even bad news immediately than to have to live with the uncertainty longer. In some decision-making tasks, people will pay to reduce uncertainty, even when knowing the outcome does not change their decision (Shafir, Simonson, & Tversky, 1993). Within personality psychology, researchers have argued that some people are better able to live with uncertainty than others and that this forms a stable personality trait (Sorrentino & Roney, 2000). In the context of our task, it may be that some people are more comfortable dealing with the uncertainty of multiple categories than others are. Others may identify the most likely category and then suppress attention to other categories, which would only make the decision more complex. People who constantly deal with uncertain situations, such as commodities traders or professional poker players, no doubt become extremely good at hedging their bets when uncertain. However, most people may have trouble dealing with the uncertainty inherent in domains such as personal finance. The economic collapse of 2008 suggests that very many people made financial decisions that were reasonable only if positive scenarios were guaranteed, in spite of consistent advice to diversify to hedge against such disasters.

Conclusion

So, do people hedge their bets in category-based induction by using multiple categories? Yes and no. At least in conscious category-based induction, people do not always use multiple categories. On the other hand, almost all subjects could use multiple categories when the task encouraged it. Anderson's (1991) proposed rule (or something like it) seems to be one option that people follow when the task encourages them to do so. When one category gives a clear, plausible answer, however, people tend to stop there and not consider other options, much like other reasoning tasks in which they focus on the most likely situation without considering other possible situations. Finally, some people strongly tend to use multiple categories in completing this task, whereas others (*more* others) strongly tend not to. The many people in the middle can be shifted to different strategies, although their overall tendency is to focus on a single category.

The goal for future research, then, is not to try to discover "the" model of category-based induction but to specify the conditions under which people do and do not consider multiple possible options. The technique we have described here provides a useful tool for that investigation.

References

- Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, 98, 409–429.
- Evans, J. S. B. (2007). *Hypothetical thinking: Dual processes in reasoning and judgment*. Hove: Psychology Press.
- Evans, J. S. B., Handley, S. J., Over, D. E., & Perham, N. (2002). Background beliefs in Bayesian influence. *Memory & Cognition*, 30, 170–190.

- Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, 19(4), 25–42.
- Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review*, 102, 684–704.
- Groopman, J. (2007). *How doctors think*. Boston: Houghton Mifflin.
- Hayes, B. K., & Newell, B. R. (2009). Induction with uncertain categories: When do people consider the category alternatives? *Memory & Cognition*, 37, 730–743.
- Heit, E. (2000). Properties of inductive reasoning. *Psychonomic Bulletin & Review*, 7, 569–592.
- Johnson-Laird, P. N. (2006). *How we reason*. Oxford, United Kingdom: Oxford University Press.
- Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, 80, 237–251.
- Kersten, D., Mamassian, P., & Yuille, A. (2004). Object perception as Bayesian inference. *Annual Review of Psychology*, 55, 271–304.
- Koehler, J. J. (1996). The base rate fallacy reconsidered: Descriptive, normative, and methodological challenges. *Behavioral and Brain Sciences*, 19, 1–53.
- 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, and 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.
- Murphy, G. L., & Ross, B. H. (2007). Use of single or multiple categories in category-based induction. In A. Feeney & E. Heit (Eds.), *Inductive reasoning: Experimental, developmental, and computational approaches* (pp. 205–225). Cambridge, England: Cambridge University Press.
- Murphy, G. L., & Ross, B. H. (in press). Category vs. object knowledge in category-based induction. *Journal of Memory and Language*.
- Papadopoulos, C., Hayes, B. K., & Newell, B. R. (2009). Non-categorical approaches to property induction with uncertain categories. In N. A. Taatgen & H. V. Rijn (Eds.), *Proceedings of the 31st annual conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.
- Ross, B. H., & Murphy, G. L. (1996). Category-based predictions: Influence of uncertainty and feature associations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22, 736–753.
- Shafir, E., Simonson, I., & Tversky, A. (1993). Reason-based choice. *Cognition*, 49, 11–36.
- Siegler, R. S. (1995). How does change occur? A microgenetic study of number conservation. *Cognitive Psychology*, 28, 225–273.
- Siegler, R. S., & Crowley, K. (1991). The microgenetic method: A direct means for studying cognitive development. *American Psychologist*, 46, 606–620.
- Sorrentino, R. M., & Roney, C. J. R. (2000). *The uncertain mind: Individual differences in facing the unknown*. Hove, England: Psychology Press.
- Stanovich, K. E. (1999). *Who is rational? Studies of individual differences in reasoning*. Mahwah, NJ: Erlbaum.
- Stanovich, K. E. (2009). *What intelligence tests miss: The psychology of rational thought*. New Haven, CT: Yale University Press.
- Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, 24, 629–640.
- Verde, M. F., Murphy, G. L., & Ross, B. H. (2005). Influence of multiple categories in inductive inference. *Memory & Cognition*, 33, 479–487.

Appendix

Models and Binomial Parameters for the Three Conditions With Probability Matching

There are four basic models that differ in whether people consider only one category or multiple categories, crossed with whether they treat features as independent or else consider feature conjunctions. For each model, we can calculate the probability that on any trial the subject chooses the multiple-category choice, p . We then can calculate q , the probability of a single-category choice. Sometimes p and q do not sum to 1 because of another (unlikely but nonzero) possibility, however, these are rarely chosen; therefore, for purposes of the model, we ignore those cases and increase p and q proportionally to sum to 1 (i.e., p is set to $p/(p + q)$). The binomial distribution then predicts the number of subjects who will give 0–4 multiple-category responses, as shown in Figure 4. Note that including the low-frequency responses would make the model fits in Figure 4 even worse, because the model would predict even fewer perfectly consistent responses. Rather than a deterministic view (i.e., always go with single- or multiple-category response), this analysis assumes that subjects are matching the probabilities given by the models. For each condi-

tion, we work out the predictions for the question discussed in the main text, as an illustration.

Condition 50/50: Figure 1—Given a Circle, What Color?

Single/independence. Dan has the most circles and an equal number of red and green, so if only this category is used, $p = q = .50$.

Single/conjunction. Dan's circles are also half red and half green, so even if one only examines circles, $p = q = .50$.

Multiple/independence. The probability of green and red are calculated in the main text, $p = .50$, $q = .33$, so with proportional responding, these are $.60 (.50/.833)$ and $.40 (.333/.833)$.

Multiple/conjunction. If one only considers circles, then 6/9 are by Dan, and 3/9 are by Preston. $P(\text{green}|\text{circle}) = P(\text{Dan}|\text{circle}) \times P(\text{green}|\text{Dan}\&\text{circle}) + P(\text{Preston}|\text{circle}) \times P(\text{green}|\text{Preston}\&\text{Circle}) = 6/9 \times 3/6 + 3/9 \times 2/3 = .556$. Similarly, $P(\text{red}|\text{circle}) = .333$. Ignoring the one purple diamond by Preston and assuming proportional responding, $p = .625$, $q = .375$.

(Appendix continues)

Condition 60/40: Figure 2—Given a Diamond, What Color?

Single/independence. Drew has the most diamonds and has 3 red (multiple-category response) and 4 blue (single-category response) figures. As calculated in the main text, $p = .375$ and $q = .5$, so with proportional responding, .4286 and .5714, respectively.

Single/conjunction. If only looking at diamonds in Drew's drawings, half are red and half blue, so $p = q = .50$.

Multiple/independence. The probability of red and blue are calculated in the main text, $p = .50$, $q = .333$, so with proportional responding, .60 (.50/.833) and .40 (.333/.833).

Multiple/conjunction. If one only considers diamonds, then 6/9 are by Drew, 3/9 by Julian. If calculated as shown in the earlier condition, $P(\text{red}|\text{diamond}) = 6/9 = .667$, $P(\text{blue}|\text{diamond}) = .333$.

Condition 80/20: Figure 3—Given a Square, What Color?

Single/independence. Federico has the most squares and has 3 purple (multiple-category response) and 4 green (single-category

response) figures. As calculated in the text, $p = .375$ and $q = .5$, so ignoring the single blue circle, with proportional responding, .4286 and .5714, respectively.

Single/conjunction. If only looking at squares in Federico's drawing, 3 are purple and 4 are green, so $p = 3/7 = .4286$, $q = .5714$.

Multiple/independence. The probability of purple and green are calculated in the text: $p = .6164$, $q = .3836$.

Multiple/conjunction. If one only consider squares, then 7/11 are by Federico, and 4/11 are by Cyrus. If calculated as shown in the earlier condition, $P(\text{purple}|\text{square}) = .6364 = p$, $P(\text{green}|\text{square}) = .3636 = q$.

Received July 17, 2009

Revision received December 7, 2009

Accepted December 9, 2009 ■

**Call for Papers: *Journal of Experimental Psychology:*
Learning, Memory, and Cognition
Special Section on Neural Mechanisms of Analogical Reasoning**

The *Journal of Experimental Psychology: Learning, Memory, and Cognition* invites submissions of manuscripts for a special section on the Neural Mechanisms of Analogical Reasoning to be compiled by Associate Editor Miriam Bassok and Guest Editors Kevin Dunbar and Keith Holyoak. The goal of the special section is to showcase high-quality research that brings together behavioral, neuropsychological, computational, and neuroimaging approaches to understanding the cognitive and neural mechanisms that are involved in analogical reasoning. The editors are seeking articles on analogy and related cognitive processes (e.g., schema induction, metaphor, role-based relational reasoning, category-based induction) that either present original research using methods of cognitive neuroscience or that present behavioral research (including studies of cognitive development and/or aging and studies of brain-damaged patients) strongly connected to the neural mechanisms of analogical reasoning.

The submission deadline is **October 1, 2010**. The main text of each manuscript, exclusive of figures, tables, references, or appendixes, should not exceed 35 double-spaced pages (approximately 7,500 words). Initial inquiries regarding the special section may be sent to Miriam Bassok (mbassok@u.washington.edu). Papers should be submitted through the journal's submission portal (see www.apa.org/pubs/journals/xlm/) with a cover letter indicating that the paper is to be considered for the special section.