THINKING & REASONING, 2012, 18 (1), 81–117

Psychology Press
Taylor & Francis Group

Reasoning with uncertain categories

Gregory L. Murphy[1], Stephanie Y. Chen[1], and Brian H. Ross[2]

[1]Department of Psychology, New York University, New York, NY, USA
[2]Department of Psychology, University of Illinois, Urbana-Champaign, IL, USA

Five experiments investigated how people use categories to make inductions about objects whose categorisation is uncertain. Normatively, they should consider all the categories the object might be in and use a weighted combination of information from all the categories: *bet-hedging*. The experiments presented people with simple, artificial categories and asked them to make an induction about a new object that was most likely in one category but possibly in another. The results showed that the majority of people focused on the most likely category in making inductions, although there was a group of consistently normative responders who used information from both categories (about 25% of our college population). Across experiments the overall pattern of results suggests that performance in the task is improved not by understanding the underlying principles of bet-hedging but by increasing the likelihood that multiple categories are in working memory at the time of the induction. We discuss implications for improving everyday inductions.

Keywords: Bayesian processes; Categories; Induction; Reasoning.

Research on categories and concepts has emphasised that induction is an essential function of concepts. Knowing that your neighbour has bought a dog allows you to predict that you will hear barking and tells you how to interpret scratching or whining you hear coming from behind the

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

We thank Marjorie Rhodes, the ConCats research group, and anonymous reviewers for helpful comments and Rebecca Bainbridge for help in running the experiments. The research was supported in part by NSF grants BCS-1128769 and 1128029.

http://www.psypress.com/tar http://dx.doi.org/10.1080/13546783.2011.650506

neighbour's door. If you are asked to take care of the dog for a day, you can predict what is going to be required of you. Without any specific information about this pet, knowing about dogs in general allows you to make predictions and interpret information about it.

Research on category-based induction takes place at the intersection of the topics of concepts and reasoning. Category-based induction is a reasoning process that takes conceptual representations as its inputs. The way that people retrieve, combine, and evaluate the information they have stored about categories is complex, and researchers have developed a number of models of the induction process (e.g., Osherson, Smith, Wilkie, López, & Shafir, 1990; Sloman, 1993). Work in our laboratory has focused on a specific reasoning problem that can arise in the induction process, namely how people deal with uncertain categories.

In the example above, when your neighbour tells you she has bought a dog you can be certain that it is in fact a dog. If your sister says she has bought a new Nissan, you probably feel sure it is a Nissan. However, in many situations you may not have enough information to classify an object with certainty. A car going by seems to be a Nissan but is possibly one of those new Hyundais; your illness is probably a cold but possibly the flu; a painting could be late impressionist or possibly fauvist in style. In such cases you can still use category information, but you must take into account your uncertainty. Normatively, the best prediction integrates answers across categories or, as we shall refer to it, *bet-hedging*. If the car may be a Nissan, with probability p, or a Hyundai, with probability q, then your prediction about its miles-per-gallon (mpg) should be the Nissan mpg times p plus the Hjundai mpg times q; that is, weighting each category by the likelihood that the car is in it. For predicting discrete properties, one cannot take the average in quite this way, but one can calculate a probability (Anderson, 1991). If you want to estimate the probability that the car is a diesel, and if you know the proportion of diesel cars each manufacturer makes, you could estimate the probability as $p \times$ proportion of Nissan diesels plus $q \times$ proportion of Hyundai diesels. Again, this weights each category's induction by its likelihood.

Anderson (1991) proposed this solution in what he described as a Bayesian analysis of induction[1] as part of his larger theory of category formation. His proposal can be seen as an application of the law of total probability (Mood, Graybill, & Boes, 1974):

$$P(\mathrm{A}) = P(\mathrm{A}|\mathrm{B})P(\mathrm{B}) + P(\mathrm{A}|\text{–}\mathrm{B})P(\text{–}\mathrm{B}),$$

[1]In all our experiments the categories are novel and equally probable, so we ignore the prior probability component of Bayesian reasoning. We continue to use the term *Bayesian* because of the common feature of Bayesian models of induction that predictions are integrated across multiple categories, weighted by their likelihood.

where A refers to the prediction and B and –B to the relevant categories (Nissan and not-Nissan, i.e., Hyundai). Although people approximately follow the rule when they are asked to give estimates of the components in some contexts (see Zhao & Osherson, 2010, and the General Discussion), they may not always do so in category-based induction.

Murphy and Ross (1994; see also Malt, Ross, & Murphy, 1995) investigated whether people used such an approach when making inductions with uncertain categories, and they found that people generally did not. In particular, varying the properties of the *less likely* category (e.g., the Hyundai) did not influence people's predictions (see also Hayes & Chen, 2008). In contrast, varying the properties of the most likely (*target*) category did affect people's predictions (Murphy & Ross, 2005; Ross & Murphy, 1996). Murphy and Ross concluded that people generally focus on the most likely category and derive a prediction based on that single category, in contrast to the bet-hedging ideal. Murphy and Ross (2010b) proposed that this focusing on a single category was related to other reasoning shortfalls in which people focus on a single possible outcome even when it would be relatively easy to take into account two or more possibilities (Stanovich, 2009). For example, Evans (2007) proposed *the singularity principle* as a general property of hypothetical reasoning, that people consider only one situation at a time, unless something prompts them to expand their thinking such as a hint or failure to derive an answer. Research on decision making suggests that people dislike uncertainty and will pay to reduce it, even when this has no effect on their ultimate decisions (Shafir, Simonson, & Tversky, 1993).

The goal of the present article is to better understand when people do and do not use multiple-categories in category-based induction. Forming a profile of circumstances that lead to single versus multiple category use will be critical for a theory of when and why people engage in Bayesian reasoning of this sort. From a practical standpoint it is important to discover what situations might increase people's accuracy in this task. Focusing on a single category when another category is also fairly likely leads to suboptimal inductions. Techniques that improve people's performance in our task may also work in major real-life predictions made in medical or career decisions.

We first briefly review a paradigm we have used in previous work and then explain what our new experiments will add to what we have learned from that work. Although focusing on a single paradigm has limitations, it also allows us to accrue knowledge across related experiments in order to arrive at a more complete conclusion.

PAST STUDIES OF CATEGORY-BASED INDUCTION UNDER UNCERTAINTY

Figure 1 shows a display based on one we used in Murphy and Ross (2010b). In this experiment participants viewed coloured figures that were

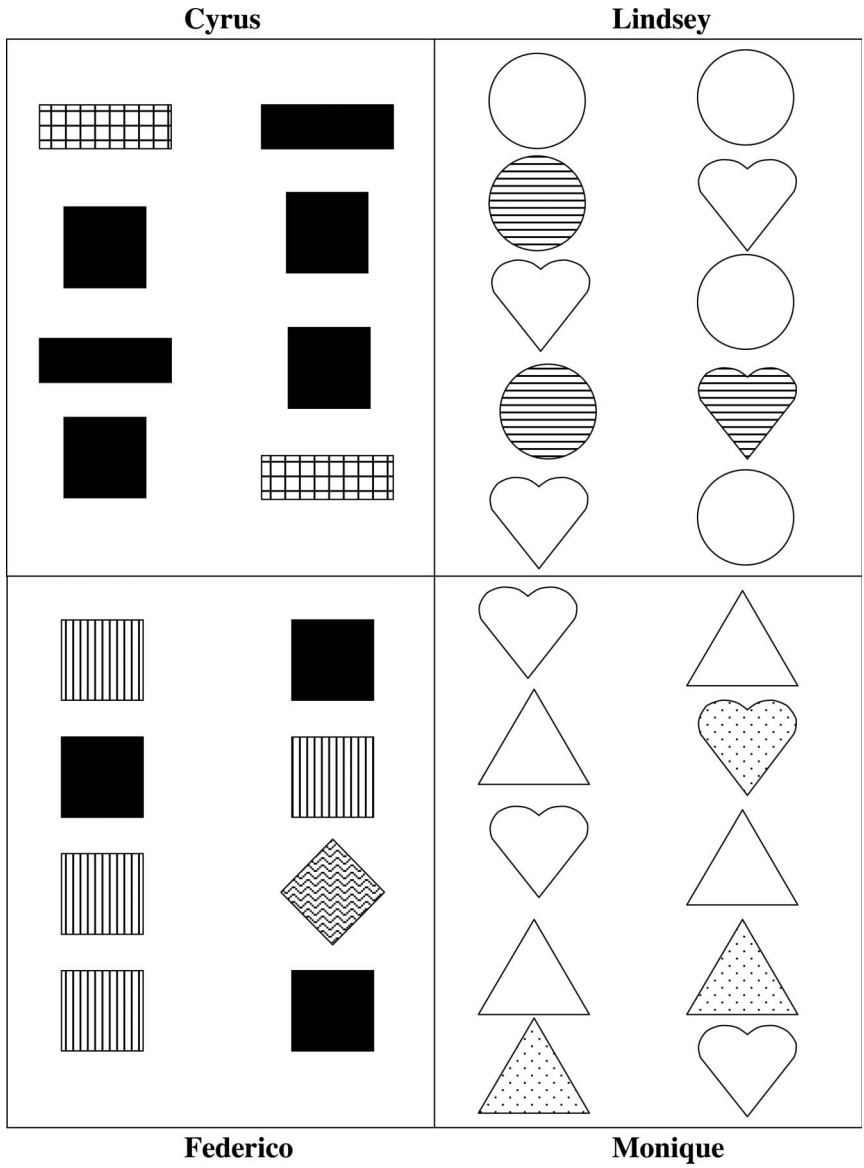

Figure 1. Combined displays from Experiments 1 and 2 (the left two categories) and Experiment 3 (all the categories). In the actual experiments the figures were coloured. The colours are represented by different patterns here. In Experiment 1, participants were asked to predict the pattern (colour) of a new square. Federico, with seven squares, would be the target category, and Cyrus, with four squares, would be the alternative. In Experiment 3, participants were also asked to predict the shape of an empty figure. Lindsey and Monique have both drawn seven empty figures, so they are equally likely.

purportedly pictures that different children drew using a computer drawing program, in which they could choose shapes and colours. (Our figures represent colours by different patterns, so we will refer to patterns instead of colour throughout this article.) It was pointed out that different children preferred different shapes and patterns. Then participants read about a *new* shape that had been found, for example a new square. They were asked which category they thought the square was in (i.e., which child drew it), the probability that it was in this category, and the pattern they thought this square would have—the induction.

In Figure 1 it can be seen that Federico was most likely to have made a square, but Cyrus also made a few squares. People generally recognised that Federico was not guaranteed to have made the square, as they rated the probability of this categorisation around 65%. If they nonetheless used only Federico's category to make the induction, since it is the most likely, they would predict that the square would be vertically striped, because Federico has four striped figures and three black ones (and the same is true for squares in particular[2]). If instead they followed the bet-hedging principle, they would have used both Federico and Cyrus to generate their prediction. In that case, they would have predicted that the square should be solid black—either by the kind of computations described above for the Nissan/Hyundai example, or by a simpler computation of counting up the colour of the displayed squares. That is, although Federico's drawings have a small advantage for striped figures, when combined with Cyrus's drawings the advantage swings to black figures.

The results showed that people chose the *multiple-category response* (i.e., solid black) only 30% of the time in such problems (Murphy & Ross, 2010b, Exp. 1). An analysis of individual participants showed that only 7 of 47 consistently followed this rule across different inductions. Instead, most answers were the *single-category response* (i.e., vertically striped), and 22 participants consistently gave such answers across different problems. This confirmed the general findings of earlier work using similar displays, as well as experiments using short stories with verbally stated uncertainty about a character or object (Malt et al., 1995; Murphy & Ross, 1994; Ross & Murphy, 1996). The results also revealed that this tendency was not monolithic, as a minority of people did seem to follow something like Anderson's (1991) Bayesian rule.

Furthermore, Murphy and Ross's (2010b) Experiment 2 found that people could be persuaded to pay attention to all the relevant categories

[2]The displays were constructed so that the predictions were the same whether people treated shape and colour as independent (as Anderson, 1991, suggests) or instead restricted their prediction to other objects with the same given feature (squares in this case). There is evidence that people do something like the latter in this task, as discussed in Experiment 5.

with a seemingly simple methodological change. In the usual paradigm, we ask people to choose the most likely category and to rate its probability. This is done in part to make sure that people agree on which category is most likely to be correct. If they chose Cyrus as most likely to have drawn a square, for example, they would likely predict that the new picture was black, but not because they were using multiple categories. Note that the probability rating is an explicit indication that participants are not certain of their choice. If they rate Federico as 65% likely to be the correct category, this can only be because they have noticed the squares that Cyrus made. Therefore, this rating explicitly requires participants to acknowledge their uncertainty. Surprisingly, that acknowledgment doesn't lead them to attend to Cyrus in the subsequent induction.

In their Experiment 2, Murphy and Ross (2010b) asked people to rate the probability of *all four* categories—not just the most likely one. So, a participant might rate Federico as 70%, Cyrus as 30%, and the other two categories as 0% likely to have drawn a square. (Most responses were of this sort.) Now when asked to predict the object's colour, 87% of the answers for this category structure were the multiple-category response—a huge increase from the 30% in the earlier experiment. Giving a probability for each category eliminated the focus on a single option, overcoming the singularity principle (Evans, 2007).

THE PRESENT RESEARCH

The current experiments investigated in more detail just what about this task makes people aware that they should be using multiple categories and why people appear unaware of this in the standard version. Why does the listing of multiple probabilities lead to a change in strategy? The simplest explanation would be that simply drawing people's attention to the less likely categories makes them include them in their computations. However, the present experiments will show that merely drawing attention to a category does not necessarily lead to its being considered, or to correct performance. Furthermore, it isn't clear that when people give the right answer they are doing so because they are implementing the principle of bet-hedging. Perhaps successful manipulations bring that principle to consciousness; or perhaps they act through less explicit channels of activating categories and bringing them into working memory without the underlying principles playing a role. Indeed, implicit measures of bringing an alternative category to mind increase the use of multiple categories (Ross & Murphy, 1996).

The experiments also investigated different procedures that might disrupt the singularity principle and cause more accurate responding. For example, if categories are completely uncertain, then perhaps people will be more

likely to attend to multiple ones. Finally, we investigated in more detail just how many people spontaneously use the multiple-category strategy. In the Murphy and Ross (2010a, 2010b) paradigm that answer can occur by chance or guessing. If people are not paying too much attention they might choose the multiple-category response even if they considered one category (or none!). We constructed a control condition that allows us to discover how often this happens, providing a better estimate of how often people use multiple categories in induction.

In summary, the present experiments investigate when and why people overcome their tendency to focus on a single category, even when they are uncertain that it is correct. The goal is to develop a profile of situations that encourage or discourage people from using multiple categories, thereby giving greater insight into the underlying causes of this non-normative reasoning strategy.

GENERAL METHODOLOGY

Participants were NYU students or other members of the community who served in the experiments for course credit or pay.

The materials consisted of displays of the sort shown in Figure 1, except that colour was used instead of the patterns shown there. People received general instructions about the children's drawing program, and they also received information about the 0–100% probability scale, both essentially identical to those used in our past work (e.g., Murphy & Ross, 1994, 2010a, 2010b). The displays were printed on paper and placed in plastic sheet-protectors in a loose-leaf binder. For each category structure, participants turned a page to see a display and then read a series of questions in a separate booklet. The details of the questions are provided later in each experiment. There were generally three questions about each display, one of them a filler, which served to disguise the purpose of the experiment. After completing that page in the booklet, the participant turned to the next set of questions and pictures. The experiments usually had four such displays and sets of questions, and they took about 15 minutes to complete. For each display one critical question involved predicting shape given the figure's colour, and the other involved predicting colour given the figure's shape.

There were always two versions of each form (except in Experiment 5) that switched the predicted features of the single-category and multiple-category strategies. That is, in Figure 1, people who attend only to Federico would predict a square to be vertically striped; those who attend to both categories would predict solid black. In a second form the striped and black patterns would be exchanged. This way, any preference for a given feature would be balanced across the two induction strategies. Half the participants received each form in each experiment.

EXPERIMENT 1

As explained above, in Murphy and Ross's (2010b) Experiment 2 people wrote down the probabilities of each category, as in (2) below, rather than choosing the most likely one and rating its probability, as in (1). This had an enormous effect on the induction rule used, as almost all inductions used multiple categories.

(1) I have a square. Which child do you think drew it?
What is the probability that the child you just named drew this?
What colour do you think the figure has?

(2) I have a square. What is the probability that each child drew it?
Federico ___% Cyrus ___% George ___% Tony ___ % (must sum to 100)
What colour do you think the figure has?

This result is very interesting, because a seemingly small manipulation overcame people's strong tendency to focus on a single category. Logically, when participants rated Federico as 65% likely to have drawn the square, they must have been attributing the remaining 35% probability to Cyrus, the only other child who drew squares. However, this logical fact is apparently different from actually writing down that Cyrus had 35% probability, as in (2). In our original question, (1), the secondary category is implicitly used in deciding the probability rating; in the newer version the secondary category is explicitly acknowledged. This seems likely to be important in explaining why people overcame the singularity principle when answering (2).

In Murphy and Ross (2010b), the questions were compared across experiments, and given the importance we will attribute to this effect in our analysis of the task, it seems wise to replicate it with participants randomly assigned to experimental forms. We compared the *standard question*, (1) above, to the *all-category question*, (2), in which people estimated the likelihood of all four categories. The only difference was that we used the question, "Which child do you think most likely drew it [the figure]?" in the standard condition rather than "Which child do you think drew it?", thereby emphasising that the participant was not indicating certainty in writing down a name. This "most likely" language was used in all the subsequent experiments.

Method

This experiment used three different displays of the "children's drawings" stimuli as described in the General Method. Each display had two critical

questions and one filler. A total of 48 participants performed the task, randomly assigned to one of the two question types just described.

Results and discussion

We first established that people selected the target category as the most likely one. For the all-category question, we accepted any trial on which the target category had the highest or tied for highest probability. One participant in the standard condition was deleted for making more than one classification error. Accuracy of choosing the target category as the most likely (or tied) was .97 in both conditions. Furthermore, the estimated probability that this category was correct was very similar as well: 69.7 and 67.7% in the standard and all-category conditions. Thus any difference in the inductions made by these groups cannot be attributed to differences in classification.

The main question is what inductions the two groups made. Referring to Figure 1, if someone predicted that a new square would be vertically striped, this would reflect a single-category induction; if he or she predicted the square would be black, this would reflect use of multiple categories. The standard group made multiple-category inductions 34% of the time, but the all-category group made them 69% of the time. Although not quite as large as the original effect reported in Murphy and Ross (2010b), the effect is still very large and is significant, $t(45) = 3.53$, $p < .001$. Nine subjects in the all-category group consistently used multiple categories across all problems, with only two consistently using single categories. In the standard group these numbers were three and eight. These results show that deriving and writing down the probability for the less likely category is effective in encouraging people to attend to it when making inductions. The slight change in wording from the earlier procedure, asking which child was "most likely", had no apparent effect.

EXPERIMENT 2

Writing down the probabilities of all the categories in Experiment 1 could be having two effects: making participants commit to the secondary category being a real possibility (a *commitment* explanation) or simply making information about the secondary category more salient (an *informational* effect). When people write down probabilities for two or more categories in question set (2) they are taking an active step to confirm, "It's not just Federico; there are other possibilities." By writing a non-zero probability for Cyrus they are committing to Cyrus being a factor in the induction. This active step could then encourage them to think about multiple categories when they get to the prediction. If the initial classification question is totally

eliminated in this paradigm, people tend to use multiple-categories more often (Hayes & Newell, 2009; for more discussion of classification effects on later judgements see Busemeyer, Wang, & Lambert-Mogiliansky, 2009), which suggests that commitment could be a critical factor.

Alternatively, once one has written that Federico and Cyrus have 65% and 35% of the likelihood, information on the questionnaire now encourages multiple category use. In version (1) of the questions there is nothing about Cyrus, and so that information remains implicit, even though Cyrus's drawings must have been consulted to derive the probability for Federico. After answering (2), the information that Cyrus has a 35% probability is salient and so will be taken into account. On the informational account, then, it is not the writing down or commitment that has an effect so much as the mere availability of a non-zero probability for Cyrus.

In the present experiment we use a technique (suggested by Hakwan Lau) in which the informational content of version (2) is maintained without any personal commitment. In this method we simply tell people the probability that a figure is in each category. For example, we might say, "I have a figure that is 65% likely to be drawn by Federico and 35% likely to be drawn by Cyrus. What colour do you think it has?" (To prevent people from deriving their own probabilities we omitted any information about the other features of the object.) In terms of informational salience, both Federico and Cyrus have been mentioned, and their (correct) probabilities provided. Therefore, if it is merely the availability of the information about both categories that is essential, people should choose the answer that uses information from both categories more often than in a control condition using the usual form shown in (1). However, if people must themselves commit to Cyrus being important, then we should not see any difference between the new informational version and the standard version.

Method

The experimental booklet contained three different displays. Each display had two critical questions and one filler; one of the critical questions in each display was relevant to this comparison (the other will be described in the control experiment below). In the *informational* version, participants were told the probabilities of the two children (categories) that were most likely to have drawn a new figure. The probabilities given were the same (rounded) as the actual probabilities derived from the display. For example, if Jordan had six and Adam three orange figures, the informational question would say, "I have a new figure that is 67% likely to have been drawn by Jordan and 33% likely to have been drawn by Adam. What shape do you think the figure

most likely has?'' Following this was a list of shapes, and participants circled one of them. In the *standard* version, participants read the given feature (''I have an orange figure . . .'') and classified it and then made a prediction, as in the standard condition of Experiment 1. We randomly assigned 15 participants each to the two different forms.

Results and discussion

In the standard version in which people classified the item into the most likely category, 47% of responses followed the multiple-category strategy (which was somewhat larger than expected; see below). However, in the informational version, when the secondary category was explicitly listed in the problem, they followed this strategy only 32% of the time. This difference was not reliable, $t(28) = 1.01$, but more importantly, people were actually *less* likely to use the secondary category when they were told it had a significant probability of being correct—opposite to the large increase when people wrote down the probabilities themselves in Experiment 1. Eight participants had no multiple-category responses in the informational version, and only three consistently used multiple categories. In the standard version these numbers were both four.

This lack of an increase was also found in a pilot study with another 30 participants. (In that study there was an error in one of the problems, so we re-ran the experiment.) Here the informational version had 39% multiple-category responses, and the standard condition had 40%. Taken together, these results strongly suggest that simply making people aware that a category has a fair probability of being correct does not cause them to use that category in making inductions if there is a more likely category.

In short, the results rule out the informational hypothesis and thereby lend credence to the commitment hypothesis. Simply having the information about the likelihood of an alternative category does not increase use of that category—generating the probability oneself seems critical. Why does commitment have this effect? Based on our earlier theorising (Murphy & Ross, 2007) we propose that when the participant identifies a category as likely and calculates its probability, the category remains in working memory when the induction is performed. Surprisingly, simply reading the category's probability doesn't have the same effect. When people read the information that a shape was 67% likely to have been drawn by Jordan and 33% by Adam, they might have thought something like, ''Jordan is the best bet, so let's look at his category.'' As a result, Adam's features did not enter into the induction computation. Of course, a minority of participants did look at both categories, but this was not changed by mentioning the likelihood of the secondary category.

CONTROL EXPERIMENT

We have been asking why people don't generally use multiple categories to make inductions. However, multiple-category use has been somewhat greater than would be expected by past work. For example, the rate of multiple-category use was 47% and 40% in the standard conditions of Experiments 2 and the pilot study. Although the majority of responses are not the normative multiple-category one, this is still a substantial rate of multiple-category use. In the equivalent condition of Murphy and Ross (2010b) the rate was 30%, and Hayes and Newell (2009) found about 30–40% multiple-category use in similar problems. However, in earlier research based on group comparisons, effect sizes were near zero (Hayes & Chen, 2008; Murphy & Ross, 1994, 2005; Ross & Murphy, 1996; Verde, Murphy, & Ross, 2005).

In the present design, some multiple-category responses might be due to chance responding or errors. Participants might have incorrectly counted the features in the target category. In constructing stimuli we have discovered the hard way that this is very possible. One cannot notice a heart hidden in the corner. Furthermore, some colour and shape names are confusable, and in between looking at the display and marking down the response one might have turned "rectangle" into the similar-sounding "triangle". Finally, as in any experiment, some participants are not paying close attention. Sloppy or haphazard responding could result in a multiple-category answer. In earlier studies using groups comparisons such errors would average out across participants, because they would be uncorrelated with condition. Here, when we identify individual responses as using multiple or single categories, random errors can increase the count of multiple-category inductions.

In a control experiment we estimated the rate of apparent multiple-category responses that could arise from error. We used the category structures of Experiment 2, except that the secondary category no longer had any of the critical features found in the primary category. For example, consider predicting the shading of a new square in Figure 1. As discussed above, Federico is the target category and Cyrus the secondary category. Although Federico has drawn more striped figures than anything else, when also considering Cyrus the prediction would be that the square is solid black. In the control experiment, we altered the secondary category so that it did not have the critical features in the target. That is, Cyrus would have no black (or striped) figures. If participants miscounted or weren't paying attention, they might nonetheless predict "black", the less likely feature. We call this the *pseudo-multiple-category response*. The rate at which people make such erroneous predictions provides a baseline that we can compare to predictions with the normal structure.

We constructed four displays similar to those used earlier. Each had one category pair of the sort tested in Experiment 1 (e.g., Cyrus-Federico in

Figure 1) and one pair representing the control condition just described. The dependent measure was how often people chose the multiple-category feature in the standard condition or the corresponding (minority) feature in the control condition. A total of 19 participants served in the within-participant design.

In the standard condition people made the multiple-category choice 35% of the time, the level found in previous experiments. The question, then, is how often people chose the equivalent (pseudo-multiple-category) feature when it was not in multiple categories (and was the minority feature in its target category). That mean was 12%. This difference was reliable, $t(18) = 2.20$, $p < .05$, showing that there is more multiple-category use than can be accounted for by error or chance. However, more significant is that fully a third of the multiple-category responses seem due to error or a strategy that does not involve multiple categories.

A follow-up experiment used the informational technique described in Experiment 2, in which we provided the probabilities of the categories rather than describing a feature. People chose the multiple-category choice 32% of the time in the standard displays (as reported in Experiment 2), compared with 10% pseudo-multiple-category responses in the controls, $t(14) = 2.1$, $p < .06$, confirming that about one-third of seemingly multiple-category responses result from "error" of some kind.

In sum, the control experiment suggests that about a third of the "multiple category" responses found in other experiments might be due to errors, and so the level of multiple-category use is less than the total response rate of the earlier experiments would suggest. Considering our own experiments (above and Murphy & Ross, 2010b) and those reported by Hayes and Newell (2009), a reasonable estimate is that around 25% of responses truly integrate across categories.

EXPERIMENT 3

We now return to the question of just why the majority of people focus on a single category. The results of Experiment 1 show that it is not particularly difficult to use both categories to make the induction, when people choose to do so. The limitation seems to be on people realising that they ought to attend to both categories, rather than being unable to effectively combine information from both (Murphy & Ross, 2007).

The reasoning literature also includes examples in which people do not consider multiple alternatives even when it is easy to do so. Stanovich (2009, p. 70) gives an example:

> Jack is looking at Anne but Anne is looking at George. Jack is married but George is not. Is a married person looking at an unmarried person?
> A) Yes B) No C) Cannot be determined.

One of the present authors will admit that he chose "Cannot be determined", even though he encountered the problem in a book on errors in reasoning: Since we are not told whether Anne is married, and since George and Jack are not looking at one another, we cannot tell whether a married person is looking at an unmarried person. The author, and apparently over 80% of the people tested in Stanovich's lab, thought that since Anne's status was unknown, nothing could be inferred. Of course there are only two possibilities here: Anne being married or unmarried. And it is quite easy to reason that, if Anne is married, then she is looking at an unmarried person; and if Anne is unmarried, then a married person is looking at her. Therefore, in fact, the answer is A—yes.

What is striking about this example is how easy it is to derive the answer once one considers the two possibilities. It is by no means beyond our computational resources, yet people do not think to do the necessary computations. Instead, the most likely response is to throw up one's hands and say that since nothing is known about Anne, nothing can be concluded. A parallel question can be asked in the context of category-based induction: If category membership becomes so uncertain that there is an equal chance that an object is in two categories, will people throw up their hands and say they can't answer, or will they now consider both categories?

There are at least two differences between the induction case and the Anne problem that might yield a different answer for the induction. First, we do not permit people to throw up their hands—they must make an induction (there is no "cannot be determined" option). Perhaps necessity will force them to consider both possibilities. A second difference is that our induction problems involve uncertainty of prediction, even within each category. Even if we know that Tony made the figure, we cannot perfectly predict its shape. In contrast, if we know that Anne is unmarried, there is a certain answer, perhaps encouraging people to seek certainty of her status.

In short, category-based induction could provide an easier task for people to avoid the singularity principle when categories are equally likely to be correct. Because answers can be averaged across categories (either in their values, or in their probabilities), the impulse to choose one or the other category may be less than in reasoning problems with a true or false answer. In contrast, the commitment hypothesis would predict that the equal probability case should still lead participants to use a single category. Even if they say that a particular category is only 50% likely, when they write down the percentage for that category and not the other, they are committing to that category being relevant and are being led away from the other.

To test these hypotheses we constructed problems like the following. In Figure 1, imagine that you had a new drawing that was empty inside. Which

child do you think most likely drew it, and what shape do you think this drawing has? In this case the chances are equal that the drawing was made by Lindsey or Monique: Each has drawn seven empty figures. If participants use both categories to predict the shape, they will most likely predict a heart: Across the two categories, hearts are the most likely figure (there are eight of them), and also the most likely empty figure (six of them). However, if participants focus on a single category they will find another figure that is more likely than hearts *within that category*. For example, if they choose Lindsey, they would find six circles (four empty circles)—more than the four hearts that Lindsey drew (three empty ones). Similarly, within Monique's drawings, there are more triangles than hearts. Thus their prediction tells us about which categories participants used for their induction: hearts suggest that both categories were used, circles that Lindsey was arbitrarily chosen, and triangles that Monique was arbitrarily chosen.

Given that participants have to examine both categories and realise that they are equally likely to be correct in order to answer the first question, this could have the effect of overcoming their tendency to use a single category. Having no way to rationally choose one category over the other could encourage people to use both categories. As a control condition we used examples like the Cyrus/Federico case in Figure 1. Here there was a clear most likely category, which should discourage people from integrating across both categories – that is, should encourage reliance on the singularity principle.

Method

There were 16 participants in a within-participant design. Each of four displays was constructed like that shown in Figure 1. Two of the categories in a display entered into the usual *standard condition*, in which one category was more likely (like Federico–Cyrus in Figure 1). (This condition also served as the control group in Experiment 2.) The other two were in the *equal probability condition* (like Lindsey–Monique). Note that the number of exemplars was 10 per category for these items–rather than 8 for the others–which was necessary to implement the design. To ensure that the multiple-category feature (hearts in Figure 1) was not chosen because of some inherent preference, two forms were used, in which that feature was switched with one of the single-category features (e.g., circles in Figure 1).

Results and discussion

In the equal probability condition, people were expected to realise that there were two equally likely categories. In fact, participants rated the probability

of their chosen category as being 50.5% on average, and three-quarters of all trials had probabilities of exactly 50%. The question is, then, whether this high level of uncertainty led participants to use multiple categories when making predictions.

The standard condition had 47% multiple-category responses. Therefore there is considerable room for higher levels of responding in the equal-probability condition. But in fact the equal-probability condition had a nearly identical rate of multiple-category responses, 48%. Four participants consistently used multiple categories in the standard condition, and three in the equal-probability condition. When participants did not choose a multiple-category response, they always gave the response appropriate to the category they had chosen in the first question. For example, if they chose Lindsey with a 50% probability (see Figure 1), they might predict that the shape was a circle, but never that it was a triangle. An examination of the responses revealed that when the two critical categories were presented side by side, people chose each about equally often. However, when one category was presented above the other, the top category was favoured by 14 of 16 participants, confirming that participants' choices were so uncertain that they were based on arbitrary differences between categories.

As described in Experiment 2, we performed a very similar pilot experiment. The pilot showed a similar pattern, with 40% multiple-category responses in the standard condition and 47% in the equal probability condition, which was not a significant difference, $t(15) = 0.86$.

In sum, even when the two categories were equally likely, over half of the time people simply chose one category and gave the answer associated with that category. Not only does this illustrate the strength of the singularity principle in induction, it also illustrates that the categories seem very important to (some) participants in this task. Although it has been suggested that people may be ignoring the categories in such displays (e.g., Papadopoulous, Hayes, & Newell, 2011), at least half the time people made a prediction that was based on a category that they arbitrarily selected (see also Lagnado & Shanks, 2003). In a way, categories can provide a basis for putting the singularity principle into practice. Rather than looking at 36 objects to predict a new object's colour or shape, focusing on a category allows people to evaluate only 10 items. The fact that a different category, which would have led to a different answer, could just as easily have been chosen does not discourage people from relying completely on one category and neglecting the other.

In some ways this is the most striking example of people's disinclination to use multiple categories that we have found, as they express complete uncertainty between two options and yet most often answer based on just one of them. The singularity principle seems to work as strongly here as in classic reasoning problems discussed by Evans

(2007) and Stanovich (2009). In everyday life, when categories are not explicitly presented, one can only imagine that the tendency to pick one category is even stronger, to avoid retrieving information about multiple categories from memory.

EXPERIMENT 4

The pattern of results so far suggests that when people themselves identify the alternative category as being important, they use multiple categories. In all other cases they use single categories the majority of the time. Surprisingly, *telling* people that the alternative has a 33% chance of being correct doesn't increase their use of multiple categories (Experiment 2), whereas when they themselves find the category to be about 33% likely, they do use multiple categories (Experiment 1; Murphy & Ross, 2010b).

According to our analysis (Murphy & Ross, 2007), this pattern of results suggests that when participants themselves specify a category as being relevant, though relatively unlikely, the category remains in their working memory during the induction process. They then integrate across this category and the target when making inductions. In contrast, when told that Jordan is 67% likely and Adam 33% likely to have made the figure, people merely think, "Jordan is most likely, so I'll look at his pictures", failing to integrate across categories.

There is a different explanation of these results, however. Perhaps when both relevant categories are brought to mind, people realise that the prediction requires integrating answers across the categories—they more or less consciously realise that bet-hedging is called for, because two categories are relevant. That is, perhaps the abstract principle becomes salient when writing down the second probability. If that is the case, then the need for bet-hedging should be equally encouraged by asking about only the less likely category. Since participants know that, say, Jordan is the most likely category, if they were to be asked a question about Adam, they should then realise that there are two relevant categories that need to be considered. In contrast, the usual procedure of letting the participant identify the most likely category does not lead to consideration of bet-hedging, because only a single category has been brought to mind. On this account, then, people's predictions are improved when the prediction question makes them think about the abstract principle of bet-hedging.

If our working memory explanation is correct, that principle plays little part in performance, and it is less clear that asking about the less likely category will be helpful. Doing so will bring the alternative category to attention, but that could have the effect of driving the most likely category *out* of attention. We have argued that the participant's commitment to the

two categories seems important to using both of them. If the participant commits only to the less likely category, he or she might ignore the more likely category.

Beyond the theoretical question of what is underlying the improved performance of asking about all categories, it may be useful for practical reasons to discover whether drawing attention to a less likely category helps induction. If it does, this simple manipulation could be used in real-world settings. Therefore Experiment 4 investigated this question.

We constructed a new category structure, illustrated in Figure 2, to explore this issue. This structure yields different answers if people are focusing on the alternative category versus integrating across the two categories (similar to that used in Experiment 3, but without equal probability of the categories). Consider the induction of predicting the shape of a new black figure drawn by one of the children shown in Figure 2. The two relevant categories now have three predominant shapes: diamonds, circles, and hearts. Tony is most likely to have drawn a black figure ($10/17 = .59$), with George less likely ($7/17 = .41$). If the person decided to focus on the most likely category, Tony, then she or he would likely have

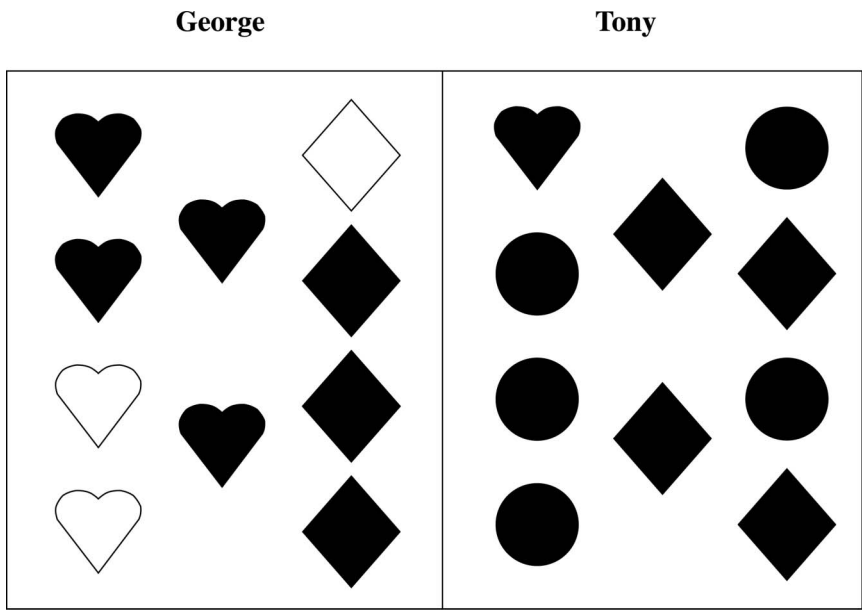

Figure 2. An example from Experiment 4 in which participants predicted the shape of a new black figure. Tony is the target category, with circle his most likely shape. George is the alternative category, with heart his most likely shape. Diamonds are most likely if the two categories are integrated. In the experiment the displays had four categories.

predicted the shape to be a circle (he has five circles and four diamonds). If the person focused on George, then hearts would win (four black hearts vs three black diamonds). However, if the person used both categories, he or she would have counted more diamonds than the number of circles or hearts. Using Anderson's formula, the probabilities are .40 for diamonds, .29 for circles, and .31 for hearts. However, it is more likely that people using multiple categories are making a simpler comparison of the shapes of black figures: seven diamonds vs five hearts or circles.

In Experiment 4, then, we constructed problems such as classifying a new black figure, followed by an induction. Half of the classification questions asked the probability that figure came from the target category (Tony in Figure 2), and half asked about the alternative category (George). Then participants predicted the object's shape. Depending on which of the three responses they gave, we could determine whether they focused on the queried category or used multiple categories in making their induction. The experimental logic is in some ways similar to that of Experiment 3 but is a more extreme test of the commitment hypothesis. Experiment 3 showed that even if the two categories have equal probability, asking participants to commit to one of them (by answering the "most likely" question) is sufficient to lead to single-category responses. Here we ask whether people will commit to the *less likely* alternative and rely on it for induction, even when they say that it is less likely.

Method

The experiment followed the usual general procedure, with 21 participants. There were three pages of displays of four categories that followed the structure shown in Figure 2 (two such pairs per display).[3] The main difference was that instead of letting participants choose the most likely category and provide its probability, the question itself mentioned a category and asked its probability; for example, "I have a new black figure. What is the probability (0–100) that Tony[George] drew this figure?" The induction question then asked the participant to predict the figure's shape. There were three questions per display. One asked the participant to estimate the chance that the *target* category child had drawn a figure (e.g., the probability that Tony drew a black figure), followed by the induction question. Another question was identical, except that it asked about an

[3]There was an error in stimulus construction such that the given feature of one pair of categories appeared as a possible predicted feature in another pair on the same display. This could have altered the inductions from the given feature, but, as the results will reveal, there was little effect given the influence of the queried category. When we removed the relevant question from the results, the means in Table 1 changed only 0–2%. Therefore we retained this question in the reported results.

alternative category in that display (e.g., that George drew a black figure), followed by the induction. However, the target and alternative questions were asked about different category pairs—the same person would not estimate the probability of both George and Tony. The questions were balanced across groups, so that each category pair was in the target and alternative conditions across participants.

A third, *filler* question asked about a category that did not have the given feature in it. In Figure 2 a filler might have asked for the probability that Tony drew an empty figure, followed by a shape prediction. The fillers were used to emphasise that merely asking about a category did not imply that the category was in fact relevant to the following induction question, so their answers were not designed to reveal how many categories were used.

We modified the usual instructions to explain that we would provide the category and that the participant was to estimate its likelihood of making the object with the given feature. "We will describe one property of an item (e.g., tell you its colour) and then ask how likely it is that one of the children drew it. Keep in mind that *this is a question*. If we ask you how likely it is that Judy was to have made the drawing, for example, we are not suggesting that it is likely (or unlikely) that she drew it. We simply want to know what your estimate is of this likelihood." Thus it was explicitly stated that merely asking about a category did not entail that it was relevant to the induction question, and the filler questions reinforced that fact.

Results

We discarded one participant who gave bizarre probability judgements to the filler categories (averaging over 80% probability for answers that should have been near 0), suggesting that she did not understand the scale or was not taking the task seriously.

The remaining participants' initial probability judgements were generally sensible, although they showed a large subadditivity effect (Tversky & Koehler, 1994). When asked about the target and alternative categories, people rated their likelihood as 80.8% and 54.4%, respectively, which total well over 100% (135.2%). This follows the well-known finding that when people focus on a particular event they are likely to overestimate its probability relative to other events. However, the target category was still rated 25% more likely on average than the alternative, so participants were fairly well calibrated in the relative likelihood of the two categories. And, in particular, even if they believed that the alternative category was 54% likely, this implies that they should use multiple categories after being asked about it.

Table 1 shows the proportions of induced features reflecting the target or alternative categories or both, as a function of the queried category. Looking first at the multiple-category responses, there is a 7% increase in

TABLE 1
Percentages of three different induction answers as a function of queried category, Experiment 4

	Predicted feature type		
Queried category	*Target*	*Alternative*	*Multiple-category*
Target	55.0	6.7	35.0
Alternative	1.7	56.7	41.7

The target and alternative features were those most frequent in their respective categories; the multiple-category feature was the most frequent across the two categories. The percentages may not sum to 100, because other responses were possible.

use of multiple categories when the alternative category was probed, but this was not significant, $t(19) = .94$. However, this does not mean that the queried category had no effect. In fact it had an enormous effect on the induction question, by shifting people from focusing on the target category to the alternative category. When the target category was queried, the most common response was the feature prevalent in the target category; when the alternative was queried, the most common response was the feature prevalent in the less likely alternative category. The shift in use of the target feature due to the queried category was significant, $t(19) = 5.29$, $p < .01$. However, there were multiple-category responses in both conditions, and the shift reflects individual differences, as we will show shortly.

As we noted in the Method, the filler items were constructed primarily to remind participants that our asking about a category did not indicate that it was particularly likely. All the fillers queried a category that did not have the given feature at all (e.g., it asked the probability that a child with no triangles had produced a triangle), so these questions were not designed to test any hypothesis. However, we noticed that some participants gave very surprising answers to these questions; namely, they produced an induction feature that was common in the queried category but not in the categories that actually contained the given feature.

To illustrate, consider Figure 2. If we asked the probability that Tony drew an empty figure, the answer should be near 0. In fact the mean rated probability for the filler questions was 2.0%. Therefore participants should have made their induction based on George's figures, since he drew the only empty figures, probably answering that the figure was most likely a heart. However, some people gave answers like "circle", which is a shape that *only* occurs in Tony's category. Having indicated that there is virtually no chance that Tony drew the figure, they then used Tony to make their induction.

To estimate how often this happened we used a conservative measure of counting only the features that were most frequent in the queried category

and that did not occur at all in the most likely category. That is, we counted how often people gave answers like "circle" when asked about Tony, ignoring diamonds or other shapes that might possibly have reflected use of George as well. A full 54% of the fillers had such predictions, in which the queried category was relied on for induction even though there was no evidence that it was relevant to the prediction.

This pattern was subject to individual differences. Ten participants always made such responses, compared to eight who never did; two were mixed. Clearly, the first set of 10 participants were greatly influenced by the queried category, even though they were told that asking a question about a category carried no information. Perhaps those participants thought that there was a chance that Tony had drawn an empty figure after all, even though he had no such figures in the display. However, their average probability rating for the filler categories was 2%, indicating that they realised that this category was not relevant—and then used it to derive the inference anyway.

We wondered if the fillers could help us identify individual patterns of reasoning in the experimental trials. In particular, when participants were asked about the alternative category ("How likely is George to have drawn a black figure?"), they could have focused only on the alternative, even though it was less likely, or they could have used both the alternative and target categories. The 10 subjects who focused on the filler categories also focused on the alternative category when it was queried in the experimental trials: They provided the alternative category feature 90% of the time. The eight participants who did not make the suboptimal response on the fillers provided the alternative category feature only 17% of the time on the test trials—they gave the multiple-category response 79% of the time. Indeed, when the question focused on the target category, these participants also gave the multiple-category response 79% of the time, which is quite high (compared to overall means of 34–47% in earlier experiments).

Discussion

The results show that there are two groups of participants. One group is highly influenced by the superficial feature of which category was queried. Even though they were told that this was not informative, and even though a third of the queried categories were obviously irrelevant to the prediction, they consistently based their inductions on the queried category. For this group, therefore, categories seem to be very important in induction. They focus on a single category and do not seem very concerned about whether it is the right one, taking commitment to an extreme.

The second group of participants seems much closer to the normative ideal. They were more likely to use multiple categories in their reasoning,

whether or not the categorisation question suggested this strategy. Furthermore, they did not use the queried category when it was clearly inappropriate to do so, unlike the first group. Interestingly, this group also had a much smaller subadditivity effect. Their rated probabilities for the target and alternative categories summed to 110%, whereas the first group's sum was 151%.

What do these results tell us about the potential explanations of why people used multiple categories in Experiment 1? If drawing people's attention to two categories brought the principle of bet-hedging to mind there, it seems likely that it should have done the same thing here. When asked about a less likely category, they would have noticed that it was less likely than the target category and then would have realised that multiple categories should be used in the induction. However, asking about the secondary category caused only a non-significant increase in multiple-category use and in fact made reasoning worse (see below). These results are inconsistent with the idea that asking people about unlikely categories brings to mind the general principle of integrating across categories. Instead the results show the negative side of the commitment effect: Simply answering a question about a category causes some people to rely on it in their inductions, even when they don't believe it is particularly likely.

What do the results say about our attempt to improve induction by mentioning the secondary category? Assuming that the correct prediction is the one that takes both categories into account, it is provided 35% of the time when the target category is queried and 42% of the time when the alternative is queried. However, this non-significant increase in accuracy must be compared to the much larger increase in the *least* likely answer—the secondary category feature. Although this is least likely by Anderson's formula and least frequent (of the three reasonable answers) in the display, choices of this answer increase from 7% to 57% when the alternative is queried. As a whole, then, asking people about a less likely category does not improve their accuracy in category-based induction.

EXPERIMENT 5

The results of Experiment 4 suggested that most people do not have knowledge of bet-hedging that is brought to consciousness by asking about multiple categories. Asking about a single less likely category did not significantly increase multiple-category use, whereas asking about all categories had a strong effect in Experiment 1. This is more consistent with the idea that questions about the categories have the effect of bringing the queried categories into working memory. A more direct test of people's knowledge of bet-hedging would be to ask them about it while they are doing this task. If people understand the basis for integrating across

categories, then they should be able to recognise this reasoning when they read it and should reject single-category reasoning.

In Experiment 5 we presented people with the usual displays and the usual categorisation questions. We then presented them with an induction question (“Which shape do you think this red figure most likely has?”) followed by a multiple-choice set of explanations of how to derive the prediction. One answer corresponded to single-category reasoning, one to multiple-category reasoning, and another to ignoring the categories entirely. Papadopoulos et al. (2011) have argued that people ignore categories and simply count the number of critical features in the entire display (e.g., count the number of red squares, red triangles, etc. to identify the shape of a red figure). Such a strategy is correlated with Anderson’s proposal, but it does not take into account categories at all. If people believe that this strategy is optimal, they should presumably choose the third option. Thus this experiment should reveal whether people recognise the validity of bet-hedging as a principle when directly asked about it.

Another set of test questions examined whether people believe that features are independent. Although not of central interest to the present investigation, this issue has been investigated in recent experiments (Murphy & Ross, 2010a; Papadopoulos et al., 2011), with the general finding that people attend to feature conjunctions. That is, when asked to predict the shape of a red figure, they tend to examine only other red figures, which makes sense only if shape and colour are assumed to be correlated. We asked participants to choose responses that assumed dependence or independence of features to discover which strategy they explicitly held. These questions also assessed whether people thought single or multiple categories were relevant.

Method

A total of 20 participants participated in the experiment. The booklets consisted of two displays like those used in the previous experiments. The first display tested people’s use of multiple categories, following the structure used in Experiments 1 and 2 (as in the Cyrus–Federico categories shown in Figure 1), and the second tested the use of correlated features as well as multiple-category use. The first two questions in each set were identical to our standard form as tested in the previous experiments. That is, the first one asked participants to identify the category most likely to be correct, and the second question asked for its probability of being correct. The third question asked for an induction of an unknown feature. The answers were written as follows, for predicting the pattern of a new square (refer to Figure 1):

A. There are more black squares than any other kind of square, so the new figure is most likely to be black.
B. Federico made more squares than anyone else, so he probably made it. He made more striped squares than black ones, so the new figure is most likely to be striped.
C. Federico made more squares than anyone, but Cyrus also made a number of them. Federico has four striped squares and three black ones, but Cyrus drew only black figures. When you put the two possibilities together, it is most likely that the new figure is black.
D. Federico made more squares than anyone. He drew a lot of black and striped figures, so he is probably sick of these patterns. Therefore a new figure of his is likely to be wavy.

Answer A is a *frequency* answer, which does not make reference to any categories. This corresponds to the strategy that Papadopoulos et al. (2011) propose. Answer B is a single-category response, in which the most likely category is identified, and its most likely shape given as the answer. (In this design the most frequent shape was also the shape that occurred most often with the given feature. The next design separated frequency from feature conjunction.) Answer C is the multiple-category response. It is longer than the others, because it must describe both categories and integrate across them. Answer D is a filler response, which was somewhat nonsensical. We varied the nature of these fillers across items, for variety. As in the previous experiments, there were two critical induction problems per display. Because the reasons given were virtually identical across questions, there did not seem to be much gained by asking the same question over and over, so we tested only one display with two critical question sets in this part. The order of options varied across questions.

The second display contained categories in which the correlations between the given feature and the predicted feature were preserved or broken. In the *preserved correlation* case the given feature occurred most often with the most frequent feature; in the *broken* case the given feature occurred most often with a less frequent feature. Murphy and Ross (2010a) discovered that the overall frequency of the predicted feature was not as important as the conjunction with the given feature. Figure 3 illustrates the structure used. One question would ask people to predict the pattern of a diamond. Here, Maura's most likely pattern, stripes, occurs for all four diamonds. So the overall category frequency and conjunction frequency coincide. In contrast, imagine being asked to predict the shape of a new dotted figure. In Anna's drawings the most likely shape is rectangles, but of the dotted drawings there are three squares and only one rectangle. In this *broken correlation* condition, then, the most frequent feature in the category

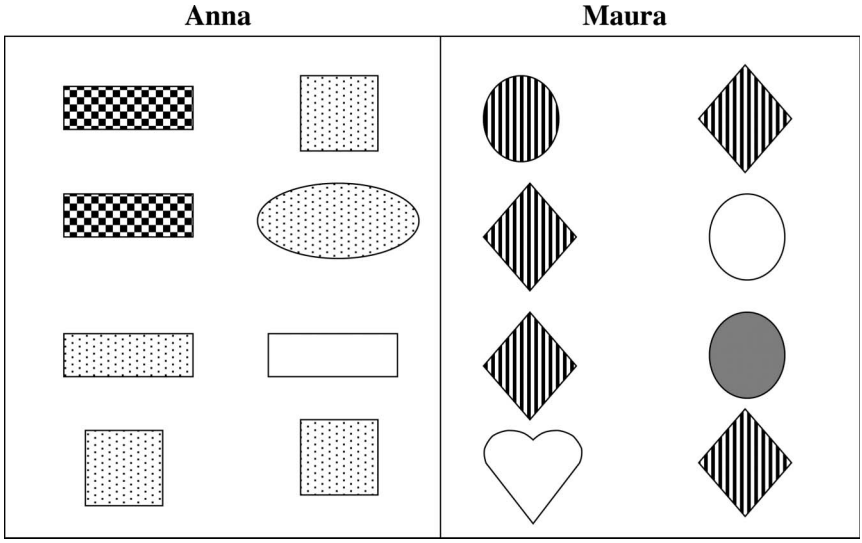

Figure 3. Part of the conjunction display from Experiment 5. One question asked participants to predict the shape of a striped figure. Maura's most common shape, diamond, is also the most frequent striped figure (a preserved correlation). Another question asked participants to predict the shape of a dotted figure. Anna's most frequent shape is rectangle, but the most frequent dotted figure is square (a broken correlation). If people are using feature conjunctions to make inductions, they should predict that a new dotted shape is a square, even though it is not as frequent as the rectangle.

was different from the conjunctive feature that occurs most with the given information.

Each induction test began with the usual categorisation and probability questions, as in (1). The subsequent induction questions offered the following alternatives for predicting the shape of a dotted figure (see Figure 3):

A. Anna has drawn more dotted figures than anyone else, so she probably made this one. Her favourite shape is rectangle, so the new figure is most likely to be a rectangle.
B. Karla [not shown in Figure 3] has drawn a dotted figure that is a heart, so the new figure is most likely to be a heart.
C. Anna has drawn more dotted figures than anyone else, so she probably made this one. Three of her dotted figures are squares, and two are different shapes, so the new figure is most likely to be a square.
D. If you look across the whole display, more dotted figures are squares than anything else, so the new figure is most likely to be a square.

A describes the *within-category independent* strategy in which the most frequent shape of the most likely category is predicted. B is a filler. C uses the *within-category conjunction* strategy. Within Anna's figures the shapes with the given pattern are counted. (Note that in this question the answer is different depending on whether one uses frequency or conjunction, rectangle vs square. This was true in two of the induction problems.) D uses the *no-category conjunction*, in which conjunctions are counted up with categories ignored.

There were three separate critical questions for the conjunction display. In addition to discovering whether people use feature conjunctions, the difference between options C and D tests for whether they use categories or ignore them, as in the first display.

Results and discussion

The data of interest are the frequencies with which people identify the different answers as being the correct form of reasoning. Across 40 responses in the first display, 18 corresponded to single-category reasoning (45%), and 14 to multiple-category reasoning (35%). Only four ignored the categories entirely, as suggested by Papadopoulos et al. (2011). What is striking about this is that even when the correct reasoning pattern of integrating across categories is described, a majority of respondents do not choose it. Contrary to both the views that people are optimal in their use of categories or that they ignore categories entirely, many people focus on a single category even when the possibility of integrating across two relevant categories is explicitly offered to them.

The second display tested people's use of feature conjunctions as well as their use of multiple categories. Consistent with the findings of Murphy and Ross (2010a), we found that people professed a within-category conjunction strategy 41 times (68%), compared to a within-category independence strategy only 14 times (23%). There was no apparent difference between the preserved and broken correlation conditions (they chose the conjunction strategy 75% and 65% of the time, respectively). People chose a less likely feature in the category if it occurred with the given feature more often (e.g., predicting from Figure 3 that a new dotted figure would be a square rather than a rectangle). Only five responses followed Papadopoulos et al.'s (2011) prediction that people use conjunctions while ignoring categories. In terms of conscious strategies at least, it seems that people believe that feature conjunctions are important for making inductions. Equally important for the present purposes, participants chose to focus on within-category conjunctions, rather than referring to shapes across all the categories. The structure of these categories had greater within-category consistency than in the other displays, which might have increased the tendency to focus on a

single category. Nonetheless, people did not claim to ignore the categories and use conjunctions when given the chance.

It is puzzling that our results seem to contrast with those of Papadopoulos et al. (2011) so strikingly. For example, their Experiment 2 found that 98% of their participants' predictions used a conjunction strategy across multiple categories, whereas our studies have never approached this level of multiple-category use (here or in our earlier work). In the present experiment only 8% of the responses explicitly affirmed a strategy of simply counting conjunctions regardless of category membership, as Papadopoulos et al. propose. Our experiments differ in a number of details, ranging from the number of categories presented (two vs four), the precise category structures tested, and the participant populations (Australian vs American university students). Without further data on which differences might be relevant we will not speculate further. However, results from both labs agree that people tend to focus on conjunctions when possible.

Overall the results provide insight into people's explicit reasoning. Consistent with our earlier results, a large number of people prefer to focus on a single category than to use multiple categories. If we add the category ignorers to those who used multiple categories, we arrive at equal numbers of responses of those who effectively used multiple categories and those who focused on single categories in the first display. This casts doubt on the idea that people understand that information should be integrated across categories but simply do not activate this knowledge under most circumstances. In fact, half or more of the participants reject this strategy when it is offered. Once again, it seems that categories are important in this task, and many people will focus on the most likely category and ignore others.

The results of the second display are perhaps less surprising, because the use of conjunctions was strongly demonstrated in past investigations. However, they also confirm the tendency to focus on a single category. These results add a useful comparison to those of the other display, because in this design the description of the option involving a single category is longer than that involving multiple categories. In the first display, multiple-category use had the longest description.

Of course, explicit reasoning is not necessarily what is going on in our previous experiments or in real-world situations where people can simply make predictions without having to give any justification. Nonetheless, the present results speak against a simple competence-performance account of suboptimal reasoners, in which people understand what they should do but are simply responding too quickly or thoughtlessly to actually do it (cf. Frederick, 2005). Giving the "right answer" as an option in Experiment 5 did not make the majority of participants realise what they should have been doing. Instead, they clung to single-category reasoning as being the best way to make inductions.

GENERAL DISCUSSION

Five experiments investigated when and why people use multiple categories when making a prediction about an object whose categorisation is uncertain. We have used a single paradigm in this paper, in which people are given samples of the categories that they may examine when answering questions. We recognise that other paradigms with different ways of presenting information and asking induction questions could lead to different results. However, in the present design, issues of memory and classification error are much less than they would likely be in everyday life, and the results are more likely to reflect reasoning processes. Furthermore, by following this single paradigm through five experiments, we have come to a clearer understanding of what is controlling people's behaviour in this task.

Multiple-category use

The first question is simply how often people follow the normative rule of considering multiple categories. Overall, we discovered such responses 30–50% of the time in our standard condition, but our control experiment suggests that about a third of those responses do not reflect the actual use of multiple categories. Furthermore, consistent with past results (Murphy & Ross, 2010b), we found that the normative responses were often concentrated in a group of people who consistently used multiple categories. For example, we identified 23% of the participants in Experiment 2 and 25% in the control experiment as consistently using multiple categories in our usual procedure. Thus the overall mean of Bayesian responding in any experiment reflects a minority of people who consistently choose the normative response averaged with the majority of people who only occasionally give such responses, some of which are likely counting errors.

The people who consistently do not use multiple categories are nonetheless able to do so. When participants first rate the likelihood of all categories, almost 70% of inductions are Bayesian (Experiment 1; 87% in Murphy & Ross, 2010b). Thus, like a number of other reasoning errors, the problem is not that the answer cannot be computed but that people do not pay attention to relevant information, instead focusing on the most salient or available category to make their response. It is striking that many people do so even when they believed it was only 50% likely to be correct (Experiment 3).

It is interesting to compare these results with those of Zhao and Osherson (2010), who asked whether people followed the law of total probability in assigning probabilities to simple and related conditional events, such as

P(Mr X buys a car), *P*(Mr X buys a car given he wins the lottery), and *P*(Mr X buys the car given that he does not win the lottery). Overall, people's estimates of the simple probability were appropriately related to their estimates of the conditional probabilities, which is perhaps inconsistent with our results that most people don't consider two likely categories in making predictions. That is, they don't consider both *P*(shape is red given it was drawn by Preston) and *P*(shape is red given it was drawn by Don) in predicting *P*(shape is red). It may well be that asking the participants to estimate all these probabilities caused them to coordinate them more accurately than in our task. In fact the best performance in our task was also when we also asked people to estimate the component probabilities (Experiment 1). Furthermore, Zhao and Osherson also found that people's predictions of the simple event were much more related to one conditional probability than the other, usually the event that was the focus of the question. This is reminiscent of our finding that people may focus on one category that is asked about or that they select as most likely. Thus some of the same psychological processes may be underlying these two rather different experiments.

Our result also seems related to a finding in a different paradigm in which people also focus on only one possibility. Van Wallendael and Hastie (1990) gave participants a murder mystery problem with four potential suspects. As participants received new evidence, they revised the probabilities that the suspects were guilty. However, about half of these revisions were non-compensatory: They increased or decreased the probability for one suspect while leaving the other probabilities unchanged (resulting in totals that did not sum to 100%). This result is in some ways the converse of ours. We discovered that many people do not use evidence from two possibilities when making a prediction; Van Wallendael and Hastie found that people draw implications of new evidence to one possibility only, even when it logically must affect multiple possibilities.

In short, there is a strong tendency for many people to answer our induction questions using a simple, suboptimal strategy. The experiments' results give us insight into what encourages people to focus on a single category even when they recognise that it may not be the correct one. We address three important variables: the framing of the question, people's understanding of the bet-hedging principle, and the factors of working memory and salience.

Framing and linguistic variables

One possible reason that people focus on a single category is that in our standard procedure we ask which category is most likely to be correct prior to the induction question. Busemeyer et al. (2009) found differences in

induction as a function of whether the categorisation is made first, arguing that categorisation leads to a difference in framing for the later question. Perhaps in the current situation people believe that the induction question is meant to be answered *assuming* that their best guess is correct, as a Gricean implicature given the structure of the questions (Hayes & Newell, 2009). This seems unlikely to us, because the question immediately preceding the induction is a confidence rating of this classification, and participants consistently rate that they are not certain the target category is correct. Therefore, if they think they are to answer the induction question based on their previous answers, they should be paying attention to the fact that they were only 65% sure that the shape was drawn by Anna, since they have just written "65%".

More direct evidence for this conclusion is found in Experiment 2. Here we told the participants that two categories were relevant, giving their probabilities. Thus any Gricean implicature would clearly suggest that both categories should be taken into account, since we took the trouble to mention those two and not the others. However, this manipulation created no increase in multiple-category use whatsoever.

Nonetheless, asking the initial question clearly has some effect, as shown in Experiment 4 (and see Hayes & Newell, 2009; and in a different paradigm Lagnado & Shanks, 2003). If it is not this Gricean implicature, then what is the mechanism behind that effect? We provide an explanation in the next section.

Deep understanding vs category activation

The minority of participants who consistently use multiple categories even under unsupportive conditions seem likely to have some conscious knowledge of the bet-hedging principle. Even when they only identify the most likely category in the standard procedure, they spontaneously look at and take into account the less likely category. An alternative interpretation of such participants is that they pay no attention to the categories at all, which gives the same answer as integrating across categories (Papadopoulos et al., 2010) in this paradigm. Rather than integrating their answer across categories, they are simply ignoring the categories as irrelevant to the question. Contrary to this explanation, very few people claimed to use such a strategy in Experiment 5; far more said that using multiple categories was correct.

Indeed, when participants chose what they believed was the right way to answer the induction question in Experiment 5, the majority response was to use only one category even though the normative response was available. We find this result surprising. In many judgment errors, when people are confronted with the supposed right answer they accept it and

experience embarrassment that they made an error (Stanovich, 1999). One source of error is people's acceptance of a simple answer that quickly comes to mind, as they fail to do the work necessary to calculate the correct answer (Frederick, 2005; Kahneman & Frederick, 2002). Providing the answer overcomes the necessity to do that calculation and should greatly increase accuracy. Although participants in Experiment 5 might have had a tendency to choose a single-category response, when confronted with the normatively correct answer we expected them to realise that this was a better answer. However, fewer than half the responses correctly identified the multiple-category choice, suggesting that this failure is not simply a shortfall in performance—some people positively reject the principle of integrating predictions over categories when uncertain. Similarly, Tversky and Kahneman (1983, p. 300) found that 65% of their participants rejected the correct explanation of a probability conjunction problem when it was offered to them—a different error in coordinating judgements about two categories.

It is puzzling that this rate of normative responding is noticeably less than the rate of multiple-category use in Experiment 1 when we simply asked people to rate the likelihood of all the categories (69%). An explanation of the Experiment 1 result could have been that deriving all the probabilities made people realise that multiple categories are relevant in this task—that is, brought to mind the bet-hedging principle. However, if the principle is something that people know and use when it is indirectly brought to mind, then surely they ought to recognise and confirm it when it is directly presented to them. Since many do not recognise bet-hedging as the correct solution, the problem people have in doing this task is probably not that they are not thinking of the principle, because when they *are* thinking of the principle, they still don't necessarily give the right answer (Experiment 5).

Our working memory account is that, when people evaluate multiple categories, the relevant categories are brought forcefully to mind and then remain in working memory during induction. Looking at the display, seeing that two categories have triangles, and calculating probabilities for both, all conspire to keep both categories in memory when the induction question is asked. In a very different paradigm we have found that when the induction property is associated to a real-world category, people take that category into account even when it is not the most likely one (Ross & Murphy, 1996). The effective variable, then, is not that writing down all the probabilities brings to mind the importance of using multiple categories but simply that when the two categories are in working memory, they are both used.

We arrived at this explanation in order to explain a pattern of results in which people generally did not use multiple categories unless some manipulation brought the less likely category to mind (Ross & Murphy,

1996). We believe that it is a useful heuristic for predicting when people will use multiple categories in induction, but it is limited in that we do not have a complete theory of just what manipulations will bring a secondary category into working memory. For example, it seems sensible that asking people to write down the probabilities of two categories serves to keep both of them active during the later induction process. It is surprising, then, that *telling* people two probabilities does not have a similar effect. Our after-the-fact explanation is that, when people derive the probabilities themselves, they make a commitment to each category being involved in the induction. When they are simply told the two categories, they focus on the most likely one.

This conclusion is supported by the results of Experiment 4, which varied which category we queried. Once again, there was a set of normative responders who were not particularly influenced by this variable and who used multiple categories the majority of the time. However, half of the participants were strongly influenced by which category was queried, to the degree that it caused them to focus on a clearly irrelevant category—commitment gone wild. This behaviour is difficult to understand. It goes far beyond what can be explained by Gricean implicature. An implicature is a default inference, and one of its signal properties is that it is *defeasible*; that is, it can be overruled by explicit statement (Grice, 1975, p. 57; Levinson, 1983, p. 114). We explicitly told participants in the instructions that the classification question did not carry any information and that the mentioned category could be irrelevant. Furthermore, we included filler questions that *were* completely irrelevant to reinforce this fact. It is very puzzling in such a context that people would use a category that has no triangles, say, to predict the colour of a new triangle, just because of an earlier question asked about the category.

Such behaviour seems to reflect a sort of mindlessness, in which some participants are strongly influenced by a mentioned category without any notion of whether the category is relevant. Murphy and Ross (2010a) found another example of mindlessness in this task, in which people attended to a *feature* that was not relevant but that had been mentioned. As we have noted (see Experiment 5), if people are asked to predict the colour of a new triangle, many will only consider the colour of triangles in the display, apparently assuming that shape and colour are not independent. Murphy and Ross found that people continue to use this strategy when the feature mentioned was not known to be the feature of the tested figure (but one that was frequent in the category as a whole) and even after a learning procedure demonstrated that the features were independent. The mentioning of ''triangle'' was sufficient to cause people to focus on triangles in the display, whether or not it made sense to do so. In Experiment 4, merely asking about a category made some participants focus on that category even though it was clearly not relevant.

These kinds of effects are consistent with the claim that most people are more influenced by the presence of information in working memory than by abstract principles of induction or bet-hedging. Indeed, these cases of mindlessness suggest that the abstract principles play no role whatsoever for some people. When a feature and a category are mentioned, the reasoner goes to work, looking for items that have that feature in the named category, and the actual question asked or the relevance of this information do not affect the person's reasoning. In some situations such an approach will work fairly well. Assuming that the category and feature mentioned are in fact relevant to the induction, people may come up with the right answer without much effort. We have designed our experiments so that the target and alternative categories suggest different answers, in order to discover when people are considering both of them. If the real world is friendlier, then people's overall error rate may not be very high. However, when the world is not as friendly, as when the suggested category is irrelevant or when other categories give a very different answer, many people don't seem able to recover.

Improving induction with uncertain categories

What, then, can be done to improve category-based induction in everyday life? It is disappointing that some of our simple, obvious attempts did not work well. For example, simply mentioning the two most likely possibilities did not make people more likely to use them both. Furthermore, asking people only about the likelihood of the alternative category did not help either. Although this slightly increased the use of multiple categories, it had a much larger effect of encouraging people to focus solely on the less likely category, which led to the worst answer.

What did work well was asking people to identify and evaluate the most relevant categories, as in Experiment 1. Thus, this is a technique that could be used in some real-world situations, especially those in which people can identify the categories on their own. However, it is difficult to see how to adapt this technique to situations in which people may not know the categories or their properties. For example, in making medical predictions, patients will not generally know the possible diagnoses and their probabilities from their own knowledge. When making financial decisions, people may not be aware of potential investment classes and their risks and benefits. In such cases professionals will generally outline the possibilities (categories), their likelihoods, and their consequences. Unfortunately this seems very similar to the procedure in which we told people the probabilities of the categories (Experiment 2), and that did not lead to a higher rate of multiple-category use. Instead, most people seemed to reason that it was best to focus on the most likely category and ignore the other one.

One possibility would be to take a two-pronged approach: education followed by induction. For example, in the case of people making medical decisions, they could first receive materials or instruction on the different conditions that they might have (e.g., benign tumour, limited cancerous tumour, cancer that has spread throughout the organ, etc.) and what the outcomes might be of different treatments for each. In a subsequent session they could be asked to write down these possible conditions and their outcomes, and then to choose a treatment that maximises their overall outcome. By writing down the different conditions themselves, people might avoid errors of focusing on the best or worst possible outcomes alone. For example, perhaps chemotherapy has a good likelihood of success for all diagnoses, but surgery would only be helpful for one. Indeed, encouraging doctors to generate and test multiple diagnoses has been suggested as a cure for *diagnosis momentum*, the tendency to focus on a single likely diagnosis to the exclusion of others. As Croskerry (2003, p. 777) describes it, "what might have started as a possibility gathers increasing momentum until it becomes definite, and all other possibilities are excluded." Other research suggests that expertise provides protection against focusing on a single category, although only in the domain of expertise (Hayes & Chen, 2008), so perhaps such cases are the exception rather than the rule.

One way of approaching this issue would be to study in more detail the individual differences: Who is successful in solving these problems normatively? The individual differences could bring to light cognitive style or other processing differences that could serve to improve reasoning under uncertainty as a whole.

Finally, we note that our experiments all involve explicit, presumably conscious reasoning about categories. In important decisions people are very likely to consciously consider the options and think about what might happen if one or the other one occurs. Certainly people consider possible options when planning for college, careers, investments, and other important decisions. However, when people make speeded predictions based on an uncertain stimulus, they may paradoxically perform better. They may take into account multiple categories precisely because higher-level cognitive processes do not engage to focus attention on a single category and exclude other possibilities (Chen, Ross, & Murphy, 2010; Newell, Paton, Hayes, & Griffiths, 2010; Verde et al., 2005). In combination with our demonstrations that most people *can* consider multiple categories in explicit reasoning (Experiment 1), this suggests that improving predictions is possible, if the problem is framed in a supportive way.

Manuscript received 27 June 2011
Revised manuscript received 4 November 2011

REFERENCES

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

Busemeyer, J. R., Wang, Z., & Lambert-Mogiliansky, A. (2009). Empirical comparison of Markov and quantum models of decision making. *Journal of Mathematical Psychology*, *53*, 423–433.

Chen, S. Y., Ross, B. H., & Murphy, G. L. (2010). *Category-based induction in action and thought*. Poster presented at the 51st Annual Meeting of the Psychonomic Society.

Croskerry, P. (2003). The importance of cognitive errors in diagnosis and strategies to minimize them. *Academic Medicine*, *78*, 775–780.

Evans, J. S. B. (2007). *Hypothetical thinking: Dual processes in reasoning and judgement*. Hove, UK: Psychology Press.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, *19*(4), 25–42.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics 3: Speech acts* (pp. 41–58). New York: Academic Press.

Hayes, B. K., & Chen, T-H. J. (2008). Clinical expertise and reasoning with uncertain categories. *Psychonomic Bulletin & Review*, *15*, 1002–1007.

Hayes, B. K., & Newell, B. R. (2009). Induction with uncertain categories: When do people consider the category alternatives? *Memory & Cognition*, *37*, 730–743.

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 49–81). Cambridge, UK: Cambridge University Press.

Lagnado, D. A., & Shanks, D. R. (2003). The influence of hierarchy on probability judgments. *Cognition*, *89*, 157–178.

Levinson, S. C. (1983). *Pragmatics*. Cambridge, UK: Cambridge University Press.

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

Mood, A. M., Graybill, F. A., & Boes, D. C. (1974). *Introduction to the theory of statistics* (3rd ed.). New York: McGraw-Hill.

Murphy, G. L., & Ross, B. H. (1994). Predictions from uncertain categorisations. *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, UK: Cambridge University Press.

Murphy, G. L., & Ross, B. H. (2010a). Category vs. object knowledge in category-based induction. *Journal of Memory and Language*, *63*, 1–17. doi: 10.1016/j.jml.2009.12.002

Murphy, G. L., & Ross, B. H. (2010b). Uncertainty in category-based induction: When do people integrate across categories? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*, 263–276. doi: 10.1037/a0018685

Newell, B. R., Paton, H., Hayes, B. K., & Griffiths, O. (2010). Speeded induction under uncertainty: The influence of multiple categories and feature conjunctions. *Psychonomic Bulletin & Review*, *17*, 869–874.

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

Papadopoulos, C., Hayes, B. K., & Newell, B. R. (2011). Non-categorical approaches to feature prediction with uncertain categories. *Memory & Cognition*, *39*, 304–318.

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.

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

Stanovich, K. E. (1999). *Who is rational? Studies of individual differences in reasoning*. Mahwah, NJ: Lawrence Erlbaum Associates Inc.

Stanovich, K. E. (2009). *What intelligence tests miss: The psychology of rational thought*. New Haven, CT: Yale University Press.

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

Tversky, A., & Koehler, D. J. (1994). Support theory: A nonextensional representation of subjective probability. *Psychological Review*, *101*, 547–567.

Van Wallendael, L. R., & Hastie, R. (1990). Tracing the footsteps of Sherlock Holmes: Cognitive representations of hypothesis testing. *Memory & Cognition*, *18*, 240–250.

Verde, M. F., Murphy, G. L., & Ross, B. H. (2005). Influence of multiple categories in inductive inference. *Memory & Cognition*, *33*, 479–487.

Zhao, J., & Osherson, D. (2010). Updating beliefs in light of uncertain evidence: Descriptive assessment of Jeffrey's rule. *Thinking & Reasoning*, *16*, 288–307.