Available online at www.sciencedirect.com
SCIENCE DIRECT
COGNITION
Cognition 95 (2005) 175–200
www.elsevier.com/locate/COGNIT

The two faces of typicality in category-based induction☆

Gregory L. Murphy[a,*], Brian H. Ross[b]

[a]Department of Psychology, New York University, 6 Washington Place, New York, NY 10003, USA
[b]Beckman Institute, University of Illinois, 405 N. Mathews Ave., Urbana, IL 61801, USA

Received 17 November 2003; accepted 15 January 2004

Abstract

Studies of category-based induction using different methods have found somewhat contradictory results for whether typical items are a stronger basis for induction. Typical category items are generally more similar to other category items than are atypical ones, and they are also more likely to be categorized into the category in question. We propose that the first aspect (representativeness) influences induction, but the second (uncertainty about the correct category) does not. Two experiments using artificial categories found support for this prediction. Two further experiments manipulated pictures of objects and also found that representativeness in the category influenced the strength of induction, but uncertainty of classification did not. Thus, the two aspects of typicality have different effects on category-based induction.

Keywords: Induction; Categorization; Typicality; Representativeness; Judgment

A central function of categories is their ability to tell us properties of novel entities. When going to a new restaurant, we know what to expect and what to do; when ordering the chicken, we have some idea of how it will taste and what its nutritional qualities will be; when picking up the fork and knife, we know what they are to be used for. Even though we may never have been in the restaurant before and almost certainly have never

☆ This research was supported by NIMH grant MH41704. We would like to thank Lewis Bott, Bob Rehder, and the reviewers for their helpful comments.
* Corresponding author.
E-mail address: gregory.murphy@nyu.edu (G.L. Murphy).

0022-2860/$ - see front matter
doi:10.1016/j.cognition.2004.01.009

encountered that chicken or that silverware before, we can use our knowledge of the categories they are members of to make fairly accurate predictions about them. Thus, category-based induction is an important part of intelligent behavior.

In the restaurant example just described, you would probably be quite certain in identifying entities as spoons or chicken. But how should induction proceed when categorization is not certain? Since the work of Rosch in the 1970s (e.g. Rosch, 1973), it has been recognized that people are not sure of the categorization of some objects. Uncertainty of classification can arise in two ways. First, the uncertainty may be intrinsic to the object, as less typical items are less likely to be classified into the category (e.g. McCloskey & Glucksberg, 1978; Murphy & Brownell, 1985; Smith, Balzano, & Walker, 1978). Some people will identify a sink as a kitchen utensil, and some will not (Hampton, 1979). Second, sometimes the cause is limited information about the object, such as if it is not seen or heard very clearly (at night or at a distance). Thus, even though you would normally agree that Rover is a dog, if seen behind the bushes at night, you might not be sure.

What kinds of inductions should one make when one is not certain about the category an object is in? As a general rule, it seems that when one is not certain about the object's categorization, one should also be less certain about any induction that is based on that category. If you cannot tell for sure that the animal behind the bushes is a dog, then you shouldn't be so likely to attribute dog properties to it. Of course, this also depends on what other categories it might be in and on the properties you are inducing. For example, if you are trying to decide whether this animal breathes (we won't ask why), then your uncertainty about its being a dog is not that important. If it is not a dog, then it is probably a cat, a raccoon, or some other mid-sized mammal, all of which breathe. Thus, your uncertainty need not affect your prediction in this case. On the other hand, if you are trying to predict whether it will engage in a dog-specific behavior such as barking, then your uncertainty in categorization should lead to uncertainty in induction.

Anderson (1991) provided a Bayesian analysis of the effect that uncertainty ought to have on inductions (or as we often call them, *predictions*). For each object containing features F and for each category k, one can predict the presence of a novel feature N using the formula

$$P(N|F) = \sum_k P(k|F) * P(N|k) \tag{1}$$

That is, one calculates for the object how likely it is to be in each category k (e.g. dog, cat, raccoon, etc.) and how likely that category is to contain the property in question (e.g. how likely dogs, cats, raccoons, etc. are to breathe, or to bark). Then one sums across all the categories in order to make the prediction. In short, this proposal is that people use multiple categories to make predictions when the categorization is uncertain. Of course, when one is 100% sure that the object is a dog, then the other categories have no effect (because $P(k|F)$ would be 0 for all other categories). The net result of this Bayesian rule is to make prediction more accurate, as one hedges one's bets by taking into account all the possible categories.

1. Tests of the Bayesian view and implications for induction

Although Anderson's proposal provides a clear normative prescription for how to make category-based predictions when the categorization is uncertain, it does not generally seem to capture how people make such predictions. Under most circumstances, people tend to use a single category to make their inductions, even if they are not at all certain that that category is correct. We have found such results using a situation in which an object is most likely—but not at all certain—to be in a *target category*. We then manipulated the properties of the next-most likely category (the *alternative category*) to see whether it influences people's judgments. Under most circumstances, it does not (Malt, Ross, & Murphy, 1995; Murphy & Ross, 1994). Even if one takes some trouble to emphasize that the initial categorization is uncertain (Ross & Murphy, 1996), subjects do not change their confidence in their predictions. (There are some exceptions to this generalization, which are described in Murphy & Ross, 1999; and Ross & Murphy, 1996.)

As a general rule, then, people seem to make category-based inductions based on only one category, even when they are not certain that the object is in that category. For example, even if they give a fairly low rating of their confidence in the category, this does not lead them to use multiple categories in making their predictions. Apparently, the difficulty or uncertainty that people have in classifying an object is then forgotten (or treated as irrelevant) when making predictions about the object. Although Anderson's (1991) proposal for taking uncertainty into account would increase accuracy of induction, people do not generally incorporate categorization uncertainty into their predictions.

These findings suggest an interesting possibility, that typicality may not lead to changes in people's induction. Atypical items, as described earlier, are less certain to be in a category. They are rated lower, and more people judge them as not being category members (Rosch, 1973). Furthermore, the same person may change his or her mind across trials in deciding whether atypical items are in the category (McCloskey & Glucksberg, 1978). Thus, if categorization uncertainty does not affect induction, then typicality may not either. This conclusion would be an important and interesting fact about category-based induction. However, there is good reason to believe that this conclusion is not true.

Many studies of category-based induction have followed a paradigm in which properties are described of one or more category members, and then judgments are made about whether that property is also true of another category member. We call this the *Rips paradigm*, after Rips (1975). For example, Osherson, Smith, Wilkie, López, and Shafir (1990) used problems such as the following:

Robins use serotonin as a neurotransmitter
Therefore, sparrows use serotonin as a neurotransmitter

Other arguments relate the properties of one item to the category as a whole:

Robins use serotonin as a neurotransmitter
Therefore, all birds use serotonin as a neurotransmitter

In Osherson et al.'s experiment, subjects compared arguments that differed only in the typicality of the premise and judged which one was stronger. They found (see their Table 1), that typical category members yielded much stronger inductions from a category to its superordinate (e.g. the inference from robin to all birds was stronger than from penguins to all birds). Using a similar technique, Rips (1975) found that typical items led to stronger inductions in examples like the first one above, from one category to another at the same level. In summary, typical items are a stronger basis for induction than are atypical items (see Heit's, 2000, review for further examples).

There are many differences between the Rips paradigm and the ones we have used in our research, so direct comparisons are not possible (though see General Discussion). We used an instance that is of uncertain categorization, and subjects had to decide which category or categories to use in making a prediction about that instance. The Rips technique is to draw inductions across categories, and the premises themselves are generally entire categories rather than a single item. Nonetheless, it is troubling that typicality has a strong effect on induction in one paradigm but categorization certainty, which is highly correlated with typicality, has no effect in the other paradigm.

2. The nature of typicality and a possible resolution

We believe that the key to resolving this seeming contradiction is a more detailed analysis of typicality. As typicality is a major component of conceptual structure, clarifying its role will illuminate our understanding of concepts. Rosch and Mervis (1975) proposed two related determinants of what makes an item typical. The first is that typical items tend to have properties that are found in other category members. A robin is medium-sized, has wings, flies, lives in a nest in a tree, lays eggs, chirps, and so on. In contrast, a penguin lays eggs but lacks the other common bird properties. Thus, it is no surprise that robin is much more typical than penguins. We call this difference *representativeness*, referring to how similar an item is to other items in the same category.

Rosch and Mervis (1975) also found that atypical items tend to have the properties of other related categories. For example, a penguin swims and can spend considerable time under water, it walks, and it eats fish. These properties are more common in categories other than birds. Rosch and Mervis showed experimentally that both of these variables—sharing features with other category members and sharing features with contrast categories—influenced category learning and categorization.

We propose that the apparent difference in results between the two paradigms we have described arises from the difference in these two components of typicality. In our previous research, we manipulated nontarget categories when categorization was uncertain. For example, in one experiment (Murphy and Ross, 1994, Experiment 1), subjects rated their certainty of categorization at about 55%. We varied the constitution of alternative categories, while the properties of the target category (and therefore the item's representativeness) were held constant. In contrast, the Rips paradigm almost always has involved items in which there is little or no doubt about the categorization of the item. For example, Osherson et al. (1990) tested typicality by comparing robin to penguin and mouse to bat (see their Phenomenon 8). However, these college-student subjects did not really doubt that penguins are birds or that bats are mammals. Instead, what varied was the representativeness of the stimulus: Penguins do not fly, and robins do, and therefore the first is atypical of birds, and the second is not.

The theory that Osherson et al. (1990) propose for category-based induction also relies purely on category representativeness. In their model, the similarity of the premise category to the relevant category is one of two determinants of induction strength. The uncertainty of categorization is not part of the theory. Indeed, their stimuli were all familiar categories whose membership in the tested categories is well known. Given such items, uncertainty of categorization is probably not an issue.

In contrast, Murphy and Ross (1994) tested individual instances in unfamiliar categories, whose category membership was somewhat uncertain; Malt et al. (1995) tested individual items in familiar categories, but the items lacked complete information, and so they could have been in two or more categories (e.g., a person might have been a real estate agent or a repair person). Because we focused on the certainty question, we did not manipulate the similarity of the item to the target category and therefore did not vary representativeness.

In short, the two paradigms have focused on different manipulations, each related to different components of typicality, and this may be why they have yielded apparently contradictory results. This explanation is speculative, because there has been no comparison of the two manipulations, representativeness and certainty, within a similar testing situation. Therefore, there is no direct evidence that the two do in fact differ in the way that we have proposed. Given the importance of typicality to all aspects of conceptual processing (see Murphy, 2002, chapter 2) and to induction in particular, this apparent contradiction is important to resolve. That is the goal of the present article.

3. Structural indicators of representativeness and certainty

Studies of typicality have often used mathematical measures of category structure to carry out typicality manipulations. Two relevant ones are *cue validity* and *category validity* (Medin, 1983; Murphy, 1982; Rosch & Mervis, 1975; Wisniewski, 1995). Cue validity refers to the probability of an item being in a category, given that it has a cue or cues (i.e., various features F): $P(X \text{ is in Category} | X \text{ has } F)$. One can calculate the cue validity of a particular property (e.g., how diagnostic having wings is for being a bird) or of an item, defined across its set of features (e.g., how likely X is to be a bird, given that X has wings, flies, and eats moths). The cue validity of a feature is simply the number of items in the category having the feature divided by the total number of items that have the feature. Thus, for a feature that occurs at least sometimes in a category, the simplest way to manipulate cue validity is to vary the number of objects outside the category that have that feature. By increasing or decreasing the denominator of the cue validity fraction, one can change a feature's validity without changing its representativeness in the category itself.[1] Thus, we used cue validity to manipulate certainty of categorization. And as can be seen in

[1] Of course, one can also vary cue validity in some situations by changing the frequency of the feature in the category (but not if the feature occurs only in the category). However, this would also affect representativeness, as will be seen.

Eq. (1), cue validity is the first term in the Bayesian induction calculation, indicating the normative role that uncertainty is expected to have in induction.

Category validity refers to the probability that an item has a feature, given that it is in the category: $P(X \text{ has } F|X \text{ is in Category})$. Although it appears similar to cue validity, category validity can be quite different, as it depends only on the composition of the target category, and not on the presence or absence of the feature in other categories. That is, the higher the proportion of category members having *F*, the higher its category validity, regardless of *F*'s distribution outside the category. Thus, category validity is a measure of representativeness.

We used the technique of Murphy and Ross (1994) in order to allow a careful comparison of certainty of categorization (as indexed by cue validity) and representativeness (as indexed by category validity). By using artificial stimuli and categories, we could manipulate the two components of typicality independently, which is quite difficult using real categories and objects (see Experiment 3). Our expectation was that certainty would not have any effect on induction, in keeping with our earlier work. However, we expected that representativeness would affect induction, analogous to the findings of Rips (1975) and Osherson et al. (1990), albeit using a different paradigm and very different stimuli. If these predictions are borne out, then the influence of typicality on induction must be more nuanced than past discussions have proposed. Since category members differ systematically in typicality (even when categories are supposedly well defined—Armstrong, Gleitman, & Gleitman, 1983), this variable could influence virtually every case of category-based induction.

4. Experiment 1

The experimental procedure was closely modeled after that of Murphy and Ross (1994). Subjects viewed a depiction of four categories, said to be geometric pictures made by four children. Each item in the categories was a geometric shape with one of four colors and a shading pattern. Subjects were permitted to view these displays while answering all the questions, as induction was of interest, not category learning or memory.

One pair of displays was constructed to compare predictions that differed in their cue validity. Fig. 1 shows one of the displays. In one critical question, subjects were asked who they thought had drawn a square and what shading they believed it had. As can be seen, Liz has drawn more squares than the other children, and so she was the target category. Most of Liz's drawings are spotted, so this was the most typical answer to the question about shading. Subjects rated their confidence in both these questions (who drew it and what shading). Note, however, that there are a number of squares not drawn by Liz, and so subjects should not be perfectly confident that she drew this picture.

Consider a different question, about who drew a heart and what shading it has. Clearly, Rachel was most likely to draw a heart; there are few hearts drawn by other children. Most of Rachel's drawings are empty. The probability that the heart was drawn by Rachel, based on the evidence in the display, would be $6/8 = .75$, whereas the probability that the square was drawn by Liz would only be $6/12 = .50$. This, then, is the cue validity manipulation. The frequency and distribution of the induction features were equated in this comparison.

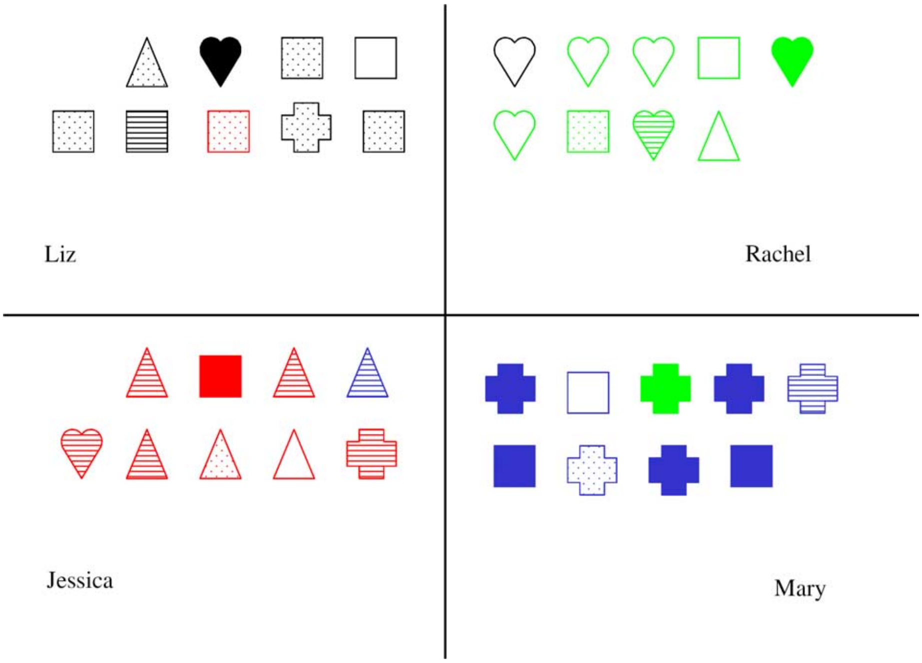

Fig. 1. Sample display used in the cue validity comparison of Experiments 1 and 2. In this display, predicting the shading of a heart was the high cue validity condition, and predicting the shading of a square was the low cue validity condition. Another display reversed these assignments.

That is, there were six dotted figures drawn by Liz and three drawn by other children; there were six empty figures in Rachel's drawings and three drawn by others. Thus, any difference in predictions of shading between the two conditions could not be due to the predicted features. In addition to the display shown in Fig. 1, there was another one in which the features' cue validities were reversed (square-dotted was high, and heart-empty was low).

It should be noted that our previous work, which also focused on category uncertainty, did not use this cue validity manipulation. Instead, it created a situation in which the correct categorization was uncertain and then varied the properties of the other categories (Malt et al., 1995; Murphy & Ross, 1994; Ross & Murphy, 1996). Thus, this experiment provides a novel test of our claim that people do not generally use multiple categories in category-based induction, by directly varying the cue validity of the target category.

Fig. 2 shows one of the displays used to test the category validity hypothesis. One question asked subjects to predict the shading of a red triangle. As can be seen, Jessica drew most of the triangles, and also all of her drawings were red. Thus, she is the target category. Furthermore, because all of her figures were red, a red triangle is very representative of her drawings. In contrast, consider the prediction about a blue cross. Mary is the target category, having drawn most of the crosses and blue items. However, note that only six of her nine figures were blue. Thus, by varying the consistency of the color in the four categories, we were able to vary the representativeness (category validity)

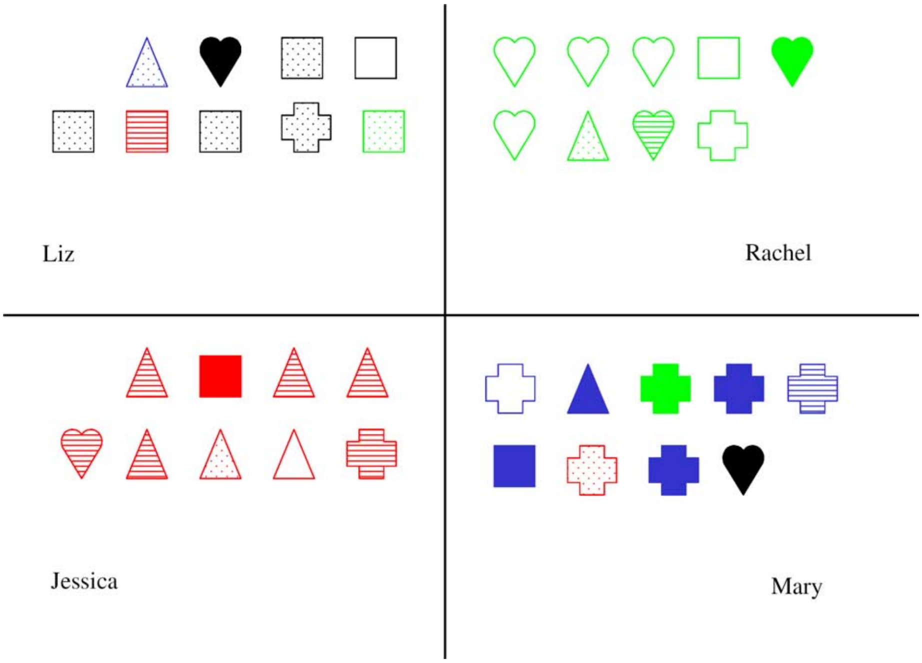

Fig. 2. Sample display used in the category validity comparison of Experiments 1 and 2. In this display, predicting the shading of a red triangle and a green heart served as the high category validity condition, and predicting the shading of a blue cross and a black square served as the low category validity condition. Another display reversed these assignments.

of the tested figure. (Note that, unlike the cue validity manipulation, category validity was manipulated within the target category.) Red and green figures were highly representative of Jessica and Rachel, whereas blue and black figures were less representative of Liz and Mary. However, because only one blue or black figure was drawn by anyone except Liz and Mary, respectively, there was little uncertainty about who had drawn the figure. That is, differences in cue validity were not an issue (and see results). In addition to the display shown, another display counterbalanced the stimuli and conditions so that blue and black had high category validity.

The goal of Experiment 1, then, was to use very similar stimuli in an identical testing situation to allow us to carefully compare the effect of cue validity and category validity in feature prediction.

4.1. Method

4.1.1. Subjects

Subjects were University of Illinois undergraduates who were randomly assigned to the cue validity and category validity comparisons. There were 31 subjects in the cue validity and 34 in the category validity conditions. (Different subjects were used in the two conditions so that a strategy used for one kind of problem would not carry over to the other).

4.1.2. Materials

The displays and their design were described above (see Figs. 1 and 2). They were printed in color and given to subjects to consult throughout the experiment. The cue validity questionnaire consisted of four sets of questions, two of which were fillers. Here is an example of one set:

I have a drawing of a square. Which child do you think drew it?
What is the probability (0–100) that the child you just named drew this?
What shading do you think the square has (circle one)
empty solid spotted striped
What is the probability (0–100) that this drawing has that shading?

Subjects circled one of the choices or wrote their responses underneath the question. The category validity questionnaire had the same form except that the drawing was described via two features (e.g. *I have a drawing of a red square*).

The effect of cue validity manipulations can be calculated mathematically, using the Bayesian formula expressed by Anderson (1991). To predict a given feature, multiply the probability that the item is in each category (given its features) by the probability that an item in that category has that feature. The sum of these probabilities is the feature prediction. So, to calculate the prediction of spotted shading given that a figure is a square (Fig. 1), one would calculate the probability that each child drew a square times the probability that this child would draw a dotted figure:

Liz: $6/12 \times 6/9 = .33$
Rachel: $2/12 \times 1/9 = .02$
Jessica: $1/12 \times 1/9 = .01$
Mary: $3/12 \times 1/9 = .03$

The resulting sum is .39. We call this the *low cue validity* condition.

One would calculate the prediction of empty shading given a heart as follows:

Liz: $1/8 \times 1/9 = .01$
Rachel: $6/8 \times 6/9 = .5$
Jessica: $1/8 \times 1/9 = .01$
Mary: $0 \times 1/9 = 0$

The resulting sum is .52. We call this the *high cue validity* condition. Each cue validity category set had one high and one low condition.[2]

[2] We could not find a way to construct a cue validity structure in which all four categories are used as target categories. The brief reason is that when the cue validity of one category is varied, one must alter the properties of the other categories, which affects their cue validities. The constraints on varying four cue validities while maintaining equality in other respects proved too difficult to resolve. In contrast, category validity can easily be varied in all four categories, because it involves only within-category properties, and so each category can be varied independently.

Note that the target category has six examples of the critical feature in both conditions. The difference between the conditions lies in the presence of the given feature (square or heart) in *other* categories, which raises or lowers the uncertainty that the given figure is in the target category.

In contrast, the category validity comparison, shown in Fig. 2, involved the target category itself. Each question in this test provided two properties (color and shape). The *high category validity* condition had uniform color within the target category, whereas the *low category validity* condition had only six of the nine figures with the target color. In this comparison, all four of the presented categories could be used in the comparison. That is, in Fig. 1, Liz and Mary were in the low category validity condition, whereas Jessica and Rachel were in the high category validity condition. In the alternate form, these assignments were reversed.

4.1.3. Procedure

Subjects were randomly assigned to one of the two comparisons (category or cue validity). They read identical instructions that explained that they were going to see examples of children's drawings and that they should study them to try to learn which children drew which drawings. After studying the categories for about a minute, they read more instructions, which explained that they would be given partial information about a figure that one of the four children had drawn and that they would be asked which child they thought had drawn it and which of the other features it had. They were told to estimate the probability that their decision was correct on a 0–100% scale, in which "0 means that something is impossible (would never happen), and 100% means that it is completely certain (would always happen in this situation). 50% means that the thing would happen about half the time." The instructions explicitly listed the features (colors, shapes and patterns) so that subjects would be sure to encode them when examining the figures. They filled out the four-question forms at their own speed, with the display available throughout the procedure.

4.2. Results

The results only are interpretable when subjects made a correct initial categorization. That is, the manipulations were made under the assumption that squares were most likely to be drawn by Liz. If a subject chose a different category, this would first suggest that the subject might not have been taking the task seriously, but it also would not allow a comparison of the cue or category validity manipulations, which were constructed based on a given target category. Thus, any response in which the incorrect category was chosen was deleted from analysis. In the cue validity comparison, because there were only two critical questions, this deletion caused the entire subject to be eliminated from analysis of inductions, because all the data from one condition would be missing.

4.2.1. Cue validity comparison

The critical results involve the probability judgment regarding the questioned feature. Across the 22 subjects who correctly categorized the items and provided the target features, the high cue validity condition had an estimated probability of 72.7%, compared to 69.0% in the low cue validity condition. This difference was not significant, $t(21) = .89$. Only 10 of the 22 subjects had a higher confidence in the high cue validity condition. Thus, this finding confirms our earlier conclusion (Murphy & Ross, 1994) that category uncertainty does not influence people's inductions—this time using a direct manipulation of cue validity.

One might be concerned about this result from two related perspectives. First, the lack of a significant result might simply indicate lack of power, perhaps resulting from variability in using the probability scale. Secondly, the difference in cue validity might not have been strong enough to generate a reliable difference (though see Murphy & Ross, 1994, Experiment 4). Cue validity would only have an effect if subjects were more uncertain about the correct categorization of the item in the low than in the high cue validity condition. As probabilities were also obtained for the initial categorization decision, we can evaluate this possibility, which turns out to address the first concern as well. All subjects who made correct categorizations were included in this analysis.

Overall, subjects provided 15.2% higher probability ratings for their categorizations in the high cue validity condition (78.6%) than in the low cue validity condition (63.3%), which was a reliable difference, $t(25) = 3.42$, $P < .01$. This result addresses both of the concerns just raised. First, the subjects could clearly use the probability scale in this paradigm to generate reliable differences. Second, subjects did observe a difference between the two conditions—they did realize that categorization was more certain when cue validity was high. However, as in many past experiments, they failed to use this difference in their own predictions.

4.2.2. Category validity comparison

As no subject was missing both trials from the same condition, no subjects were deleted in these analyses. Again, the critical data involve subjects' confidence in their predictions of the critical feature. The high category validity condition was rated 71.6%, in contrast to 63.9% for the low category validity condition. The difference between these conditions was significant, $t(28) = 2.42$, $P < .02$. Thus, in contrast to the cue validity comparison, subjects altered their confidence in predictions based on their category validity.

One might worry that the category validity effect could be a cue validity effect in disguise. Perhaps highly representative items led to stronger predictions because people were more confident that these items are in fact in the target category. For example, in Fig. 2, perhaps subjects felt that it was more likely that a red triangle was drawn by Jessica than that Mary drew a blue cross. Therefore, we did the same analysis of initial categorization decisions that we performed in the cue validity comparison. However, here there was no difference between conditions: Subjects were highly confident in their classification of both high (90.5%) and low (89.1%) category validity conditions, which were not different from one another, $t(28) = .74$. Given this high and equal confidence in initial classification, cue validity differences could not have explained the induction results.

4.3. Discussion

The two manipulations differed starkly in their effects. In the cue validity comparison, subjects showed differences in their initial categorizations, but these did not translate into differences in their inductions. In the category validity comparison, subjects were equally confident in their initial categorizations but nonetheless did show reliable differences in their inductions. In short, our hypothesis about the role of typicality in category-based induction was supported. When typicality refers to the representativeness of an exemplar in a category, it has an effect on predictions about unknown properties of that exemplar. When typicality refers to the uncertainty of whether the object is in the category, it has little effect.

5. Experiment 2

Although the results of Experiment 1 conformed to our expectations, the numerical difference between the cue validity and category validity comparisons was not very great. The cue validity effect was a nonsignificant 3.6%, and the category validity effect was a reliable 7.7%. These effect sizes are clearly not very different from one another, and it seemed wise to attempt to replicate the result, to ensure that neither the null result of cue validity nor the positive effect of category validity was a fluke. Experiment 2, then, was a simple replication of Experiment 1 in a new subject population.

5.1. Method

Experiment 2 used the same materials and procedures as in Experiment 1. However, subjects were 40 New York University undergraduates, who either received course credit or were paid for their participation.

5.2. Results

5.2.1. Cue validity comparison

Two subjects were eliminated from the prediction analyses, due to errors (see Experiment 1), leaving 18 participants. The critical data are the probability estimates of the feature prediction. Subjects rated the low cue validity condition 54.9% and the high cue validity condition 58.9%, which was not a significant difference, $t(17)=1.38$. In fact, the 4% difference between the conditions is virtually identical to that found in Experiment 1 (3.6%). In this experiment, only 5 of 18 subjects showed higher confidence in the high cue validity condition, so the small difference is clearly due to a minority of subjects. Once again, we performed the same comparison on the initial categorization judgments. The high cue validity condition was judged 10.8% higher than the low cue validity condition (72.6 vs. 61.8%), a significant difference, $t(18)=3.70$, $P<.002$. Thus, the null result of feature prediction does not reflect any problem with sample size, the probability scale, and so on, nor does it reflect a lack of difference in subjects' confidence in the categorizations. Subjects were more confident in their categorizations in the high cue validity condition, but this simply did not affect their later predictions.

5.2.2. Category validity comparison

One subject was removed due to errors leading to an empty cell. The predictions were higher in the high category validity case than in the low category validity case

(70.6 vs. 62.4%), $t(18)=2.29$, $P<.05$. Thus, category validity again had an effect on feature predictions. In order to rule out the possibility that the effect was due to perceived differences in the initial categorization (i.e., cue validity), we examined the probability judgments for the category decisions. They were again quite high and virtually identical in the two conditions (86.4 and 85.3% in the high and low category validity conditions), $t(18)=.49$. Thus, the finding of a category validity effect cannot be explained by any perceived difference in cue validity.

5.3. Discussion

The similarity in the results of the two experiments is striking. In both, cue validity had no reliable effect on induction but did have a reliable effect on categorization confidence. In both, category validity had a reliable effect on induction but no effect on initial categorization. The cue validity finding is a conceptual replication of our earlier work showing that alternative categories had little influence on induction (Malt et al., 1995; Murphy & Ross, 1994). Unlike that work, the present experiments directly manipulated cue validity. So, across the two sets of studies, we have found that manipulating the certainty of the item's category has no effect, and changing the content of the alternative categories has no effect. Thus, it seems clear that subjects are focusing on the single target category in their induction, rather than considering multiple categories or hedging their inductions due to their uncertainty. That is, they act as if their categorization is uncertain in making inductions—even though their categorization ratings reveal that they are not in fact uncertain.

The new result in this paradigm is that representativeness does have an effect on induction. This serves as a revealing contrast to the null effect of cue validity. However, it also extends the effects found in the Rips paradigm, where subjects are asked about entire categories and subcategories (like birds and sparrows). Our results concern the typicality of an individual item rather than one of these subcategories. Although typicality normally reflects both representativeness and certainty of categorization, our results show that only the former influences category-based induction.

One possible concern is that the effect of representativeness is smaller than might be expected from the past literature. Indeed, Osherson et al. (1990) found a preference of 90% for typical over atypical premise categories, which seems quite a bit stronger an effect than the present difference of about 8% in Experiments 1 and 2. Of course, it is impossible to compare these results directly, given that the tasks themselves differed, and there is no obvious way to compare typicality across the very different stimuli (is a penguin less typical than the less representative red triangle in our studies?). However, we suspect that the smaller effect that we have found is more representative of everyday category-based induction. In the Osherson et al. study, subjects examined two arguments that differed only in that a typical category member was used in one and an atypical one in the other (robin and penguin). Under these circumstances, in which subjects must select one or the other argument, the effect of typicality is likely to be large. Indeed, there is no other basis on which to make a decision, and so it is perhaps not surprising that 73 out of 80 responses favored the typical argument. In our study, which solicited ratings of individual inferences, subjects were free to pay as much or as little attention to typicality as they liked.

Support for this proposal may be found in the earlier study of Rips (1975), who presented pairs of birds and asked subjects to estimate the proportion of target birds that would have a disease given that a premise bird had the disease. Here there was no forced choice, and subjects could again pay attention to typicality as much or as little as they liked, and in fact, the raw correlation of induction with typicality of the given bird was fairly low, $r = .22$. In a second study looking at mammals, the raw correlation of typicality and induction was a moderate .48. Again, these results are not strictly comparable either to ours or to Osherson et al.'s, but the modest correlation when the items differ in a number of respects (not just typicality) suggests that representativeness has a significant but not overpowering effect on induction of individual items, as in our own results.

6. Experiment 3

Our past research has extensively used the sorts of stimuli employed in Experiments 1 and 2. However, one might well question whether results found with these very artificial stimuli would translate to real objects and familiar categories. We have examined the importance of categorization uncertainty in more realistic situations in which subjects read passages and make judgments about people who are members of familiar categories such as real estate agent and burglar (Malt et al., 1995; Ross & Murphy, 1996). There we also found that uncertainty about classification has little apparent effect on feature prediction.

In order to manipulate cue and category validity with real categories, we needed to be able to independently vary the certainty with which people think something is in a category and its representativeness in that category. The problem is that these two variables are normally closely related. For example, as items are rated less typical of a category, they have fewer of the features associated with the category (Rosch & Mervis, 1975), but they are also less likely to be judged as category members (McCloskey & Glucksberg, 1978). People are uncertain whether tomatoes are a fruit, and tomatoes also lack some of the normal properties of fruit (grown on trees, sweet taste, eaten as a snack, etc.) and so are not representative fruit. Thus, it is not possible to manipulate these variables independently by comparing different stimuli, some having low certainty of categorization and others being unrepresentative.

In order to separate these two factors as much as possible, we used two different techniques to manipulate category and cue validity. Our stimuli were manipulated pictures of familiar objects. To vary cue validity (certainty of classification), we used a cover story in which different possible categorizations of an object were provided. Depending on condition, there were either one or two plausible categories in this list. In the first case, there was low uncertainty, and in the second, uncertainty was greater. If our previous results generalize to this new paradigm, then this uncertainty should have little effect on predicting new properties of an object.

To manipulate category validity, we used a morphing program that manipulated the pictures. Such programs gradually transform a given image into another image. By stopping the program at different points along this transformation, one can obtain an image that is closer to one or the other original, or that is in the middle. Each of our stimuli was derived from two related pictures, for example, a crow and a robin, or an onion and a head of garlic. We morphed the two pictures together such that the resulting stimulus had only a small amount or a larger amount of the nontarget category. For example, the picture might be (loosely speaking) 85% crow and 15% robin or 60% crow and 40% robin. Obviously, the first picture would be more representative of a crow than the second picture would be. Thus, subjects might be more likely to attribute other crow properties (e.g. caws, eats carrion) to the first bird than to the second.

Using these two techniques, the cover story and the morphing procedure, we developed a crossed design in which a given item could have a fairly certain or an uncertain categorization, crossed with being representative or nonrepresentative of the target category. It was important to keep these variables as independent as possible, so that an effect of one could not influence the other. Under normal circumstances, a bird that is only 60% crow might not be identified as a crow by all subjects—that is, its categorization would be uncertain. Hence, this manipulation would affect not only representativeness, as intended, but also categorization certainty. We addressed this issue in two ways. First, the cover story removed the nontarget category in some conditions, so that the only plausible category was the target (crow). Thus, there was little uncertainty here. Second, as in Experiments 1 and 2, we asked subjects to classify the items, so that we could be sure that they were identifying the picture as a member of the target category. Obviously, if they identified a putative crow as a robin, then their inductions would be greatly influenced, but not because they were considering multiple categories in their prediction. Such responses were dropped from analysis (as in Experiments 1 and 2). Thus, although morphing would normally affect category certainty, the experimental design allowed us to greatly reduce this effect.

In summary, this experiment parametrically manipulated categorization certainty and representativeness. Subjects made feature predictions about the items, and their confidence was measured. If the previous results hold in this situation, only representativeness should influence induction.

6.1. Method

6.1.1. Subjects

Twenty-eight New York University students participated to fulfill a class requirement.

6.1.2. Materials

Ten pairs of categories were used, listed in Table 1. For each pair, a feature that was invisible in the pictures served as the induction property. This feature was generally true of the target category but not the alternative. For example, for the peach–apple pair, the feature was “a single, large pit,” which is generally true of peaches but not of apples.

Color photographs of similar items in the two categories were derived from the internet and various picture books. The program EasyMorph was used to morph one of the pictures into the other. By stopping the program at different steps along the way, we could derive pictures that were clearly peaches, but with a little apple in them (shown in the shininess and color), as well as pictures that were fairly ambiguous between peaches and apples. Through considerable trial and error and informal questioning of informants, we arrived at two stimuli for each pair. One was intended to be clearly in the target category although

Table 1
Categories used in Experiment 3

Target category	Alternative category	Induced property
peach	apple	has a single, large pit
orange	lemon	tastes sweet
crow	robin	eats carrion
dog	cat	would guard your home if domesticated
horse	moose	could be trained to carry riders
shark	trout	would bite a person
blueberry	strawberry	grows on a bush
rose	daisy	has thorns
duck	swan	makes a quacking sound
onion	garlic	forms rings when sliced

with some properties of the other category, and the other was intended to be somewhat ambiguous although more likely to be in the target category than in the alternative. The results will show that the pictures did differ strongly in perceived typicality. An initial practice item familiarized subjects with the morphed stimuli and the questions.

The questionnaire was similar to the one used in Experiments 1 and 2. It asked subjects which category the item was in and then it asked two induction questions. The question listed either three or four categories, as explained in the next section. One induction question involved the critical feature; the other was a filler item that was designed to prevent subjects from noticing that the questions always distinguished the two most likely categories. The filler questions were sometimes fairly true of both or of neither category. The induction questions were of the form "What is the probability that this fruit has a single, large pit?"

6.1.3. Procedure

Subjects were told to imagine that they were exploring an abandoned island in order to do a survey of its flora and fauna. They were told that some of the objects on the island were familiar but some were not. While on the island, they were able to send samples of the objects, pieces of fur, leaves, etc. to a lab for analysis. This lab does a quick analysis that suggests what the most likely categories of these objects are, rather than narrowing it down to the single correct category.

Subjects were told to first look at the picture of the object and then read in the questionnaire what categories the lab had identified as being possible for this item. These items were said to be listed in order of likelihood. In the *certain* condition, three categories were listed, and one of them was the only plausible category. For example, for the peach–apple item, the categories were peach, cherry, and raspberry. The picture looked nothing like a cherry or raspberry, so subjects should have been certain that it was a peach. In the *uncertain* condition, the alternative category was added to this list—i.e. peach, apple, cherry, raspberry. Thus, subjects would be less certain that the item was a peach, since it did look a bit like an apple. Subjects circled the category that they believed was most likely to be correct and then answered the two induction questions.

The items were randomly ordered, and two forms were used so that items that were in the certain condition in one form were uncertain in the other. Two different picture sets were used in order to vary representativeness, and subjects were assigned in equal numbers to the four resulting conditions. Each subject responded to a given item in only one condition.

6.2. Results

Trials in which subjects chose a nontarget category were eliminated from further analysis. The manipulations were performed relative to a given target category, and if subjects thought the item was in a different category, then the manipulations would not have applied. There were no empty cells in three of the four conditions, but in the uncertain/less representative condition, 9 (of 28) subjects chose the wrong category for each of their trials. (This shows that the ambiguous pictures were perhaps *too* ambiguous. However, although this means that the data from this cell are not reliable, it also means that we clearly achieved large manipulations of both certainty and representativeness, as described below.) The large amount of missing data in this cell precluded performing a two-way analysis of variance.

Table 2 presents the mean induction ratings for each condition. However, care must be exercised in interpreting the uncertain/less representative cell, as not all subjects contributed data to this cell (as just described), but all did contribute data to each of the other cells. Overall, the pattern seems to be one in which certainty (cue validity) has no regular effect, but representativeness (category validity) does. To examine this pattern, we performed *t*-tests on the three cells that did not suffer from too much missing data.

First, we compared low vs. high certainty when representativeness was high (i.e. the right column of Table 2). The difference of 2.6 was not reliable, $t(27) = .55$. Thus, cue validity did not have an effect on induction. Second, we compared low vs. high representativeness when certainty was high (i.e. the bottom row of Table 2). The difference of 9% was reliable, $t(27) = 2.98$, $P < .01$. Thus, like our first two experiments, but with real categories and a very different procedure, category validity but not cue validity seems to influence feature prediction.

One concern about the results involving cue validity is again whether the manipulation was strong enough. Here we manipulated the presence of alternative categories by the cover story, in which two or one plausible categories were mentioned. Perhaps this abstract difference did not really affect subjects' confidence in their categorizations.

Table 2
Mean probability estimates (and proportion of target categories chosen) as a function of certainty and representativeness in Experiment 3

Certainty (cue validity)	Representativeness (category validity)	
	Low	High
Uncertain	84.5[a] (.40)	85.4 (.85)
Certain	79.1 (.86)	88.0 (.98)

[a] This condition has many missing data and therefore it not strictly comparable to the others.

However, this possibility is clearly ruled out by the categorization judgments, shown in parentheses in Table 1. As the left column reveals, when category validity was low, a change in cue validity created an enormous change in categorization of 46%. Thus, the cue validity manipulation had a very strong effect on categorization. Yet, this change had no effect on feature prediction for subjects who chose the target category—in fact, the low certainty condition is slightly higher than the high certainty condition. Although there is less of a change in categorization in the right (high representativeness) column, there is still a noticeable difference, which would presumably lead to a difference in confidence of categorization. That is, when peach is the only plausible response, subjects must have been more certain that it was correct than when apple was also available, given that some subjects chose apple when it was available.

The bottom row of the table shows a possible effect of categorization certainty, as reflected in a 12% categorization difference. However, it seems unlikely that the feature prediction effect is due to any difference in certainty. First, the much larger difference in certainty shown in the first column produced no effect whatsoever. Second, the comparable categorization difference shown in the right column also had no effect. Given that the direct manipulations of certainty did not affect feature prediction, it is difficult to argue that representativeness had its effect through a smaller, indirect effect of certainty. More likely, the low representative crow, say, looks less like a crow and therefore is less likely to have crow properties attributed to it.

6.3. Discussion

The results of this experiment are consistent with the results of the first two experiments. Cue validity had no apparent effect on people's feature predictions, although it did greatly influence their categorization decisions and therefore presumably their confidence in their categorizations. In contrast, category validity did influence feature predictions. The effect is clear when cue validity was at a high level (bottom row of Table 2). When cue validity was low (top row), there is little difference. But given that one of the two cells (low certainty/low representative) is based on both fewer subjects and fewer data points per included subject (in fact, less than half the data of the other cell), it is difficult to rely much on that mean.[3]

What is interesting here is that we again see that cue validity can have an important effect on categorization itself, yet when people make predictions based on those categorizations, this effect disappears. For example, given that 15% of the subjects in the low certainty/high category validity condition actually did not choose the target category, it seems very likely that other subjects were at least considering the alternative category ("Hmmm, it does look a bit like a robin, but it seems more likely to be a crow."). Although we did not obtain confidence judgments of initial categorization in this experiment, we know from the earlier experiments that these manipulations of cue validity do influence

[3] Actually, a potentially more serious problem is that different items tended to be included in the different cells. That is, some items (e.g., the duck–swan) were more likely to have missing data than the others. Therefore, the induction features are not held constant across cells when the low/low cell is included. This problem does not arise in comparisons among the other three cells, as they had many fewer missing data.

subjects' certainty in their classification. The categorization differences suggest that the same thing is occurring here. Yet, once subjects have chosen their most likely category, any uncertainty they experienced is not used in deciding their certainty of the induction.

In contrast, a crow that is mixed with an appreciable amount of robin or a peach that has more than the usual amount of apple is not attributed the features of crows or peaches as strongly. Even if the only plausible category is the target (the high certainty condition, in which the alternative category is not provided as a possibility), resemblance to the target category still has an effect on prediction.

7. Experiment 4

One possible concern about the results is the use of an initial categorization question. Perhaps asking the subject to first identify the category suggests that the feature prediction should be performed relative to that category. If so, it is not surprising that subjects do not take into account multiple categories. More specifically for our experiments, a concern is that when subjects first categorize the item, they may then interpret subsequent questions as assuming the prior categorization. That is, they interpret the questions as meaning, "Given that Rachel drew the figure, what color do you think it has?" or "Given that it is a peach, would it be likely to have a large pit?" Thus, our null results for cue validity might reflect the fact that the initial categorization question changed people's interpretations of the induction question so that only one category was perceived to be relevant.

This possibility was addressed in earlier work (Murphy & Ross, 1994, Experiments 2 and 5), which found that eliminating the categorization question did not cause multiple category use. We therefore retained this question in the present experiments because it allows us to exclude responses when subjects choose the wrong category, which tend to create spurious signs of multiple category use. (When subjects who chose different categories are mixed together, the pattern of results is indistinguishable from that of individual subjects who consider multiple categories.) Also, note that the question immediately preceding the feature prediction in Experiments 1 and 2 asked subjects how confident they were in their categorization. Given that subjects tended to give higher confidences in the high than the low cue validity conditions, this could have created a demand characteristic to provide higher prediction ratings in the high cue validity condition as well. (This issue was not present in our earlier studies in which cue validity itself was not varied.) So, rather than biasing subjects towards considering only a single category, it is possible that our procedure biased subjects to give higher responses in the high cue validity case.

Although our previous work has not found any effect of the initial question, a recent study by Lagnado and Shanks (2003) did find a striking effect of asking people to categorize an item. However, that effect was found when no information whatsoever was provided about the category member. Nonetheless, we decided to investigate whether our results reflect the initial selection of a category. In everyday life, if you are looking at an animal and deciding whether it is friendly, you don't have to explicitly say first whether you think it is a coyote or a dog. Perhaps in the absence of this overt categorization response, you are more likely to consider multiple categories than in our task in which people made a category judgment first.

In Experiment 4, we repeated the high representativeness condition of Experiment 3, but we moved the categorization question from before the induction questions to after them. We could not entirely eliminate the question because of the sizeable proportion of subjects who did not choose the target category (see Table 2). The inductions of such subjects must be removed in order to test the hypothesis of multiple category use. However, by placing the categorization question at the end, this removed the pragmatic implication that the induction questions were relative to the earlier category that had been identified. Thus, if it is this strategy that is responsible for single category use, then we ought now to find an effect of cue validity. As this hypothesis does not concern the representativeness variable, we tested only the high representativeness condition, which was the one that provided useable data in Experiment 3.

7.1. Method

The method was similar in essence to that of Experiment 3, except that the categorization question ("Which of the lab's choices do you think is most likely the right one?") appeared after the two induction questions (one critical, one filler). Note that the form of the question emphasizes the fact that category identity is uncertain. Also, the pictures were all the high representativeness versions, which increased the probability that subjects would in fact select the target category. Cue validity was varied, as before, by listing two or only one plausible category for the item (see Experiment 3 Method). The subjects were 20 members of the New York University community.

7.2. Results and discussion

Because each subject provided data for only two conditions, rather than four as in Experiment 3, and because we used only highly representative pictures, we had more data per condition for each subject, and no one was eliminated for empty cells. The present results replicated those of Experiment 3 (see the right column of Table 2 for comparison). Subjects rated inductions in the certain condition as 81.9, compared to 78.2 for the uncertain case, which was not a significant difference, $t(19)=1.25$, $P>.20$. In fact, the size of the difference is very similar to that of Experiment 3. Once again, there is independent evidence that the cue validity manipulation did alter people's confidence in categorization, namely their categorization choices. In the high certainty condition, subjects chose the target category 99% of the time, compared to 80% in the low certainty condition, which was significantly different, $t(19)=5.60$, $P<.001$. These figures are about the same as those found in Experiment 3 (98 vs. 85%).

These results replicate the earlier findings in showing that cue validity has a large effect on people's categorization (a 19% change in classification) but no reliable effect on induction. As the categorization question appeared after the induction questions, the present method was not subject to the alternative explanation that inductions were made relative to the previous categorization choice. Thus, these results strengthen our conclusion that induction is insensitive to categorization certainty.

8. General discussion

One way to interpret our results is to say that within-category variables influence feature induction whereas between-category variables do not. Cue validity, which is the probability that an item is in one category rather than others, had no reliable effect across four experiments. Category validity, a measure of representativeness within a category, had reliable results in all three experiments in which it was tested. Normally, typicality partakes of both of these aspects. Typical items are more certain to be classified into the target category, and typical items have more of the features common in the category. Presumably, the second property leads to the first. However, the two aspects of typicality apparently have different psychological consequences. Why should this be?

In discussing the null effect of category certainty, Ross and Murphy (1996) proposed that categorization and prediction are two separate psychological processes that people usually carry out in two separate stages. Although categorization may be difficult and uncertain, once a tentative classification has been made, the processing that went into that decision has no further effects on category-based processing. After one has decided that the fruit is probably a peach, although it looks somewhat like an apple, the considerations that went into that decision are discarded and have no further influence on induction. So, in deciding whether the object has a large pit, the fact that one was thinking a second ago that it might be an apple does not change the induction, because now it is a peach, and peaches do have large pits. Indeed, the prediction ratings shown in Table 2 are all quite high, even in the conditions in which subjects must have experienced considerable uncertainty in their categorizations.

The Bayesian rule proposed by Anderson (1991) states that prediction is based on the probability that an item is in each category given its features multiplied by the probability that something in that category would have the target property. Our results show that both of these probabilities are important, but in different processes. The cue validity of an item does influence the probability that it will be classified into a category. However, this probability is functionally separated from the second component, the judgment about whether an item in that category has the critical property—the two are not multiplied together or otherwise combined in predictions.

This explanation for why cue validity has no effect on induction would also predict that category validity does have an effect. After one has decided that the fruit is a peach, one should be sensitive to how representative it is of peaches, because one is thinking of it in terms of the peach category. That is, the second stage involves interpreting the item in terms of what one knows about the selected category, and so other categories will have little effect, but knowledge of the target category will influence one's judgment. As a result, classification is even more important than one might have thought in making predictions. Once one has decided that an object is probably a peach, other possibilities are ignored, and peachiness is the salient variable.

In our experiments, there was usually a fairly clear target category that served as the basis for the induction. However, if the target category is not obvious, then people's predictions might change radically depending on which category they choose—even if they have little reason to choose that category. This is shown in a recent experiment of Lagnado and Shanks (2003). They first taught subjects categories with different properties.

They then asked subjects to estimate the chance that a randomly described item, about which they provided no information whatsoever, was in a given category. Finally, they asked the person to make an induction about this item. They found that people's answer to the first question influenced their prediction. That is, if they guessed that the item was probably in one category, their induction was more consistent with that category than if they had guessed a different category. In fact, simply by querying different categories initially, Lagnado and Shanks could change the inductions people made about an unknown item. In this extreme case, people again did not adequately take into account their uncertainty about the category. Given that they were not actually given any information about the item, they should not have changed their answers depending on the categorization they guessed. Dealing with categorization uncertainty is apparently quite difficult.

Why do people find it so difficult to take into account their categorization uncertainty when making predictions? One possibility is that there are many possible alternative categories in most situations, and it is not clear that the improvement gained from considering them all would be worth the cognitive costs involved. Especially if the initial categorization is to be used for complex planning or decision-making, keeping track of all the possible categories and their implications for each stage of the plan could surpass the limits of working memory. See the discussion in Ross and Murphy (1996, pp. 750–752) for further thoughts on this issue.

8.1. Representativeness and induction

The influence of representativeness on induction has a number of interesting implications. One is that the prediction from a chosen target category is not as straightforward as some theories propose. In the Bayesian model equation (1), the prediction from each category relies only on the category, and the item itself has no effect: $P(N|k)$ means the probability of the novel feature given the category. The item does not appear as a term in this component. (Effects of the item in this model are included in the cue validity component, which in fact seems not to influence induction.) The results, however, show a strong influence of the representativeness of the item on the prediction, suggesting that people are assessing $P(N|k)$ *and* Item I as a member of category k. Thus, the prediction is not relying on category knowledge alone, but category knowledge as "fit" to the particular item. It is possible that the fit is simply a matter of estimating overall representativeness of the item to the category. However, it seems likely that in many cases it is not just the match of the item and category representation, but the match of those with regard to a particular feature being predicted. For example, one might imagine that the overall representativeness of the crow/robin to the crow category is sufficient for determining some general crow property, such as whether it eats carrion. But, if the question was whether the item could protect its meal from the incursions of other scavengers, then it is critical to not just consider what crows do but how the particular ways in which this item matches or mismatches representative crow features that influence its ability to do that. This influence might be quite complex, perhaps involving explanations (e.g., Sloman, 1994).

Thus, our results are consistent with recent studies that have argued that an important part of category-based induction in real categories is a reasoning process in which one attempts to discover reasons for or against an item having the tested feature. For example, Proffitt, Coley, and Medin (2000) found that experts made inductions about trees not primarily via similarity relations (as in the Osherson et al., 1990, model) but often by reasoning about ecological and causal relations among trees that might account for shared properties. Similarly, studies have revealed that induction is stronger when categories are related in a way that is relevant to the induced feature. For example, biological properties are induced more strongly when categories have a biological basis than when they have a functional or situational basis (Heit & Rubinstein, 1994; Kalish & Gelman, 1992; Nguyen & Murphy, 2003; Ross & Murphy, 1999).

These kinds of variables were obviously not present in the geometric figures we used in Experiments 1 and 2, and we did not vary feature type in Experiments 3 or 4. However, our general proposal that properties of the specific item are important to induction is consistent with the knowledge effects just mentioned. So far as we can tell, all the documented knowledge effects on induction concern relations within the given categories. Proffitt et al.'s tree experts worried about how a disease might be transmitted from willows to beeches, say; they did not worry about whether the weeping willow was a tree. Instead, their reasoning focused on discovering a relevant relation between the types of trees mentioned in the question, using their specific knowledge of willows, beeches, and so on. Although our results probably did not involve such reasoning, they also demonstrate that the properties of the individual item—beyond its category membership—are important in induction, even for simple, artificial categories.

It would be interesting to extend the present paradigm to the domains and subject populations of Proffitt et al. (2000). In particular, would tree experts use cue validity in ways that our novices did not? Perhaps expertise in a domain makes one more aware of the risk of making decisions when categorization is uncertain.

A final issue involving representativeness concerns the more widely used test of induction. In the Rips paradigm that has been so popular in studies of category-based induction, whole categories are mentioned, and subjects must induce a property from one or more of the given categories to another category. In the present paradigm, subjects evaluate a specific instance relative to a number of categories it might appear in. In spite of the differences, we believe that our conclusions are consistent with the models and results from the Rips paradigm. As Osherson et al. (1990) emphasized, when a problem asks about a number of categories, people seem to judge the answer relative to a superordinate category that includes all the given categories. For example, if the problem asks about brown bears and polar bears, people consider how typical these items are of bears in general; if the problem asks about bears, dogs, and lions, people invoke the mammal category. Our claim that subjects focus on within-category properties is consistent with these models when the "category" is understood to be this implicitly evoked superordinate. That is, an induction from penguin to bluejay is rather weak, because penguin is an atypical bird—even though the bird category was not mentioned in the problem.

Models of these induction problems (Osherson et al., 1990; Sloman, 1993) do not refer to any uncertainty subjects might have about whether penguins are in fact birds, and our results suggest that they are right not to do so. Even if subjects are not completely sure that a penguin is a bird, their uncertainty would not be likely to influence their judgment. However, the atypicality of penguin would still affect induction by virtue of its lack of similarity to other items. And this is precisely what these models use to predict induction. Interestingly, a more recent Bayesian model of induction presented by Heit (2000) has the same property. This model judges the induction of a property from one category to another by estimating the proportion of properties that the premise and conclusion categories share and do not share. With this prior knowledge, the model estimates the likelihood that a new feature will fall into the shared or unshared set of features. If this model were extended to handle a single item, it would presumably again calculate the shared properties of this item (the presented crow) and the known category (crows in general), i.e. a measure of representativeness. The model does not include the variable of uncertainty (whether the item is more or less certain to be a crow). Thus, our results confirm the approach of a number of models of category-based induction and provide constraints for future approaches.

8.2. Real-world category uncertainty

The cue validity results we have found, along with those of our earlier work, may strike some as implausible in real situations. For example, a reviewer proposed the following counterexample. If you meet someone whose categorization as a doctor is uncertain, wouldn’t you be less confident that the person knows about hemoglobin than if you were certain about her profession? How can our results be reconciled with such counterexamples?

This question asks whether certainty should influence one’s confidence, and most people would probably agree that it should. However, when faced with an individual object with an uncertain categorization, one is not comparing it to a more certain case or evaluating the role of certainty in general, but is simply trying to make a prediction. Our earlier studies in fact tested cases that are fairly similar to this example and found no evidence that people consider multiple categories. For example, in one study, subjects read a story in which a person could have been either a real estate agent or (less likely) a burglar. Their predictions showed no influence of the burglar category (Malt et al., 1995)—even when the story took special pains to point out that the categorization was not certain (Ross & Murphy, 1996). However, we suspect that if we had asked the same subjects “Given that you aren’t certain that the person is a real estate agent, how confident are you…?” they would have reduced their confidence appropriately. The problem is that most real examples do not present the uncertainty as an explicit factor to be evaluated, and people do not spontaneously consider it in their predictions when it is implicit.

However, we also believe that there are limits on the effect of uncertainty (see also Verde, Murphy, & Ross, in press). The very definition of category-based induction requires a category. If one has little or no commitment to the object’s category, then presumably category-based induction would be very weak. If we randomly picked a person from the street and asked you now to judge whether that person knows a lot about hemoglobin, we doubt that you would choose a single category and make a prediction on that basis (“probably a doctor, so yes”; “probably a retailer, so no”). And our own results have found that people can in fact can use multiple categories when they are all perceived as equally likely (Murphy & Ross, 1994, Experiment 7). The question under investigation is whether uncertainty *of a likely category* affects predictions. If there is no fairly likely category, then people probably don't use categorical information and instead answer on general base rates (as in Lagnado & Shanks's, 2003, control conditions).

Finally, the reader may have noticed that in each experiment, there is a very small effect of cue validity in the expected direction. It was not significant in four out of four experiments, and the size of the effect was always around 3%. In our past work, we have not found such a trend, indeed, often obtaining effects in the wrong direction. Although one obviously cannot draw strong conclusions from an effect that never reaches significance, we suspect that this pattern obtains because one or two subjects per experiment correctly discount their predictions when categorization is uncertain. (We do not have enough data per subjects to identify individual patterns of responses, especially given that subjects rate different items in the different conditions.) We also suspect that the reason for this slight trend in the Experiment 1 and 2 is that subjects first rated their confidence in their categorization choice. Because these ratings were lower in the uncertain than the certain condition (which was not true in earlier experiments in which we did not vary cue validity), some subjects may have realized that their inductions should be lower as well.

Clearly, any such effects were of very low magnitude and not consistent across participants. Furthermore, they were smaller and unreliable compared to the differences due to representativeness (or compared to the effects of cue validity on categorization itself). However, such hints of an effect may point towards potential situations in which people do actually use cue validity in their feature predictions, which future research could investigate.

8.3. Conclusion

The results suggest that the effect of typicality on induction has two separable components, related to the two determinants of typicality initially discovered by Rosch and Mervis (1975). First, typicality relates to the probability that something will be classified into a given category: Atypical items are less likely to be classified as category members than are typical items (Hampton, 1979; McCloskey & Glucksberg, 1978; Murphy & Brownell, 1985; Rosch, 1973). But this aspect of typicality had little or no effect on induction. In all four experiments, varying the certainty with which an item was categorized into a target item had no reliable effect on induction, though it did have a reliable effect on classification certainty itself.

The second component of typicality is the degree to which an item is similar to other items in the category, or representativeness. This component consistently influenced induction. It seems likely, then, that it is the representativeness component that causes the observed effects of typicality in the Rips paradigm (e.g., Osherson et al., 1990; Rips, 1975), just as models of its results claim (Osherson et al., 1990; Sloman, 1993). Thus, there is no contradiction between past results that found that typicality (or variables associated with it) sometimes does and sometimes does not influence category-based induction.

References

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

Armstrong, S. L., Gleitman, L. R., & Gleitman, H. (1983). What some concepts might not be. *Cognition*, *13*, 263–308.

Hampton, J. A. (1979). Polymorphous concepts in semantic memory. *Journal of Verbal Learning and Verbal Behavior*, *18*, 441–461.

Heit, E. (2000). Properties of inductive reasoning. *Psychonomic Bulletin & Review*, *7*, 569–592.

Heit, E., & Rubinstein, J. (1994). Similarity and property effects in inductive reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 411–422.

Kalish, C. W., & Gelman, S. A. (1992). On wooden pillows: Multiple classification and children's category-based inductions. *Child Development*, *63*, 1536–1557.

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

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

McCloskey, M. E., & Glucksberg, S. (1978). Natural categories: Well defined or fuzzy sets? *Memory & Cognition*, *6*, 462–472.

Medin, D. L. (1983). Structural principles of categorization. In T. Tighe & B. Shepp (Eds.), *Perception, cognition, and development: Interactional analyses*. Hillsdale, NJ: Erlbaum.

Murphy, G. L. (1982). Cue validity and levels of categorization. *Psychological Bulletin*, *91*, 174–177.

Murphy, G. L. (2002). *The big book of concepts*. Cambridge, MA: MIT Press.

Murphy, G. L., & Brownell, H. H. (1985). Category differentiation in object recognition: Typicality constraints on the basic category advantage. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*, 70–84.

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

Murphy, G. L., & Ross, B. H. (1999). Induction with cross-classified categories. *Memory & Cognition*, *27*, 1024–1041.

Nguyen, S. P., & Murphy, G. L. (2003). An apple is more than just a fruit: Cross-classification in children's concepts. *Child Development*, *74*, 1783–1806.

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

Proffitt, J. B., Coley, J. D., & Medin, D. L. (2000). Expertise and category-based induction. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*, 811–828.

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

Rosch, E., & Mervis, C. B. (1975). Family resemblance: Studies in the internal structure of categories. *Cognitive Psychology*, *7*, 573–605.

Rosch, E. H. (1973). On the internal structure of perceptual and semantic categories. In T. E. Moore (Ed.), *Cognitive development and the acquisition of language*. New York: Academic Press.

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.

Ross, B. H., & Murphy, G. L. (1999). Food for thought: Cross-classification and category organization in a complex real-world domain. *Cognitive Psychology*, *38*, 495–553.

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

Sloman, S. A. (1994). When explanations compete: The role of explanatory coherence on judgements of likelihood. *Cognition*, *52*, 1–21.

Smith, E. E., Balzano, G. J., & Walker, J. (1978). Nominal, perceptual, and semantic codes in picture categorization. In J. W. Cotton & R. L. Klatzky (Eds.), *Semantic factors in cognition*. Hillsdale, NJ: Erlbaum.

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

Wisniewski, E. J. (1995). Prior knowledge and functionally relevant features in concept learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*, 449–468.