Journal of Experimental Psychology: General
2007, Vol. 136, No. 4, 685–699

0096-3445/07/$12.00 DOI: 10.1037/0096-3445.136.4.685

Blocking in Category Learning

Lewis Bott
Cardiff University

Aaron B. Hoffman and Gregory L. Murphy
New York University

Many theories of category learning assume that learning is driven by a need to minimize classification error. When there is no classification error, therefore, learning of individual features should be negligible. The authors tested this hypothesis by conducting three category-learning experiments adapted from an associative learning blocking paradigm. Contrary to an error-driven account of learning, participants learned a wide range of information when they learned about categories, and blocking effects were difficult to obtain. Conversely, when participants learned to predict an outcome in a task with the same formal structure and materials, blocking effects were robust and followed the predictions of error-driven learning. The authors discuss their findings in relation to models of category learning and the usefulness of category knowledge in the environment.

Keywords: blocking, categorization, category learning, error-driven learning

Learning which features of an object are useful for classification is not an easy task. Objects have large numbers of features, any of which might be useful individually or in combination with the other features. To complicate matters, individual features are rarely necessary and sufficient to classify an object (Rosch, 1973), and sometimes the features themselves must be constructed during learning rather than just observed (Schyns & Rodet, 1997). One goal of category-learning studies has therefore been to discover how we learn which features are important in the classification process. In a typical category-learning experiment, participants are asked to classify novel objects into one of several experimenter-defined categories. Participants gradually learn the appropriate classification by adapting their responses to the feedback provided by the experimenter, that is, by minimizing the error in their classifications. This conception of category learning has led to the development of many successful models that explain complex cognitive phenomena using error-driven learning as the basic driving mechanism (e.g., Gluck & Bower, 1988; Kruschke, 1992; Nosofsky, Palmeri, & McKinley, 1994). In this article, we ask whether all of what people learn in a classification task can be explained by the drive to minimize classification error or whether additional learning mechanisms are used.

Models of supervised category learning typically incorporate some form of dimensional weighting to indicate the importance of different features of the environment. If one feature is useful in determining an object's classification, this feature should be used more than the other dimensions. For example, if wingspan is useful in classifying birds whereas tail length is not, wingspan should be used more than tail length. But how is this dimensional weighting learned? By far the most common approach assumes that the learning system is allocating dimensional attention in a way that minimizes the difference between the predicted classification and the true classification, or the *error* in classification. So, if using the wingspan feature reduces error more than using the tail length feature, the wingspan feature is weighed more.

Some of the earliest evidence for error-driven learning can be found in studies on *blocking* in animals (e.g., Kamin, 1969; Rescorla & Wagner, 1972) and in humans (e.g., Dickinson, Shanks, & Evenden, 1984; Kruschke & Blair, 2000; Shanks, 1985). Kamin's (1969) original paradigm involved testing rats in three learning phases. In the first phase, one group of rats was conditioned to associate a shock with another stimulus, for example, a light being turned on. In the next phase, they were presented with a compound stimulus consisting of a tone plus a light, along with a shock. Finally, the rats were tested on the tone alone in order to measure its strength of association with the shock. Kamin discovered that the association between the tone and the shock was greatly reduced compared to that of a control group that did not learn the light–shock combination in a prior learning phase.

Kamin (1969) suggested that learning of the tone–shock association in the second phase was blocked because there was no longer any error that needed correcting in the rats' performance: The shock was already perfectly predicted by the light. This error-driven explanation of learning was incorporated into the Rescorla–Wagner model of conditioning (Rescorla & Wagner, 1972) as well as other more recent models (see Pearce & Bouton, 2001, for review). Furthermore, some version of error-driven learning has been incorporated into almost every model of supervised category learning, including connectionist feature-based models (Gluck & Bower, 1988), rule-based models (Nosofsky et al., 1994), exemplar models (Kruschke, 1992), and prior knowledge models (Heit & Bott, 2000; Rehder & Murphy, 2003). (We omit discussion of unsupervised learning models until the General Discussion.) The success of these models in accounting for a large range of data suggests that error-driven learning is an important part of the classification process. Indeed, Kruschke (1993) speci-

Lewis Bott, School of Psychology, Cardiff University, Cardiff, United Kingdom; Aaron B. Hoffman and Gregory L. Murphy, Department of Psychology, New York University.

Support for this work came from National Institute of Mental Health Grant MH41704. We are grateful to Sarah Hefton for programming and collecting data.

Correspondence concerning this article should be addressed to Lewis Bott, School of Psychology, Cardiff University, Tower Building, Park Place, Cardiff CF10 3AT, United Kingdom. E-mail: bottla@cardiff.ac.uk

fied error-driven learning as one of the three basic principles accounting for the success of his model ALCOVE. The analogy from conditioning to category learning can be expressed by an example. If you already know that albatrosses have a large wingspan, you might not learn about their long tails because you can already perfectly recognize albatrosses by their wingspan. Once category members can be classified correctly, there is no error left to drive the change in associations, and so new predictors of the category are not learned.

Error-driven learning has a logic that captures our intuitions about what is important to subjects in the blocking paradigm (Hoffman & Murphy, 2006). The rats in Kamin's (1969) study had already learned to predict the occurrence of the shock by the light in the first phase of learning, so why should they learn about the tone in the second phase, when doing so could not improve performance? When people (or rats) learn real-world categories, however, they may be trying to learn more than to predict category membership. It is widely agreed that the main reason that concepts are useful is that they afford inferences and generalizations about unseen properties of objects (see J. R. Anderson, 1991; Markman, 1989; Murphy, 2002; Smith & Medin, 1981). If all we knew about birds was that they have feathers and fly, we might succeed in classifying a large number of creatures correctly but we would be unable to make further inferences with this classification ability. We could not predict the behavior, habitat, or biological properties of things we identified as birds. Clearly, when more is known about a category, more inferences can be made about its members. It may therefore be sensible for an organism to learn as much as possible about categories, rather than learning only the minimum necessary to perform classification.

Thus, although it is clear that blocking is a real phenomenon of associative learning, there is reason to doubt whether it can operate freely in natural category learning. For one thing, many real-world categories are related by causal or thematic knowledge (e.g., Ahn, 1998; Rehder, 2007). For example, the bird features *has wings*, *flies*, and *is light* are all related via knowledge about flight. The features used in blocking studies are typically not related, such as the medical symptoms used by Kruschke and Blair (2000). Numerous studies have demonstrated that the presence of a theme affects category learning. For example, participants find themed categories easier to learn than nonthematic categories (e.g., Murphy & Allopenna, 1994; Spalding & Murphy, 1996), and they learn different types of category information when a theme is present (e.g., Bott & Murphy, in press; Murphy & Wisniewski, 1989; Spalding & Murphy, 1999).

Moreover, Kaplan and Murphy's (2000) results suggest that learning is not entirely error driven when themed categories are used. Kaplan and Murphy taught one group of participants novel categories in which some but not all of the exemplar features were connected by a prior knowledge theme. A control group learned categories in which none of the features was linked by a theme. The prior knowledge group displayed better learning on the thematic features than on the nonthematic features, but surprisingly, their learning of the nonthematic features was no worse than that of the control group. If learning was entirely error driven, then improved learning on the themed features should have come at the cost of reduced learning on the nonthematic features. This was not the case, contrary to predictions of error-driven learning models such as that of Heit and Bott (2000).

Associative learning experiments typically involve exemplars with fewer features than those that describe real-world categories. For example, Kruschke and Blair (2000) presented participants with two or three medical symptoms, such as "skin rash" and "back pain," for which they had to learn an associated disease. In contrast, real-world categories likely have dozens of associated features (see discussion in Murphy, 2005). The effect of numbers of features on category learning was investigated by Hoffman and Murphy (2006, Experiment 3), who compared the learning of categories with features on four stimulus dimensions to the learning of categories with eight dimensions. Participants first learned to classify exemplars (artificial bugs) into two categories and then were tested on individual features, to see how many they had learned for each category. The categories followed a family resemblance structure in which no single feature was perfectly predictive but all features were useful to some degree. In both conditions, participants needed to learn a minimum of three features to correctly classify all of the exemplars. Since error-driven learning assumes that learning stops after the exemplars can be perfectly classified, it predicts that participants would learn equal amounts in the two conditions. Instead, more features were learned in the condition with eight dimensions than in the condition with four dimensions, and classification was equally fast in the two conditions.

Hoffman and Murphy (2006) suggested that participants behaved differently from an error-driven model because they were trying to learn more about the categories than merely classifying the exemplars. Participants were interested in learning the kind of features that occurred in the categories of bugs, such as the type of antenna or the color of the eyes, independently of the need to perform the classification. Therefore, they continued to attend to features (and therefore to learn them) even when they were classifying correctly. Knowing about these different features could be useful in future situations in which, for example, some properties are not visible or when generalization is required over features other than those that are most salient. The usefulness of knowing a concept lies in the inferences it permits, and learning as much about a category as possible maximizes those inference possibilities.

This view is given some support by evidence indicating that participants learn different category representations when they are encouraged to learn categories in ways other than classification. For example, in inference learning (e.g., A. L. Anderson, Ross, & Chin-Parker, 2002; Yamauchi & Markman, 1998, 2000), participants are provided with category membership but must supply the value of the missing feature; no classification takes place. Chin-Parker and Ross (2002) found that participants were sensitive to the feature correlations within exemplars in the inference learning procedure but not in a standard classification paradigm. Similarly, Yamauchi and Markman (1998) found that participants under inference instructions learned prototypical feature values, whereas participants under classification instructions learned exemplar information. Further support comes from studies of unsupervised learning, in which participants are not given any feedback on category membership. For example, Lassaline and Murphy (1996) found that subjects who made inferences about items were later better at identifying the family-resemblance categories than were subjects who made other judgments about the items. Love (2002) found that when participants learned the Shepard, Hovland, and

Jenkins (1961) problems under unsupervised conditions, such as rating the stimuli for pleasantness, they learned the problems in a different order than those given standard classification learning instructions. These findings illustrate that learning occurs in a variety of ways that are not restricted to minimizing the error in classification.

The aim of this study was to examine an analogue of blocking in a category-learning task, as blocking is the prototypical phenomenon of error-driven learning. An error-driven learning account predicts that blocking would occur with real-world categories just as it would with feature associations: If a feature exists that perfectly classifies all objects within a category, little or no information should be learned about other features that are not perfectly predictive. This is because once objects can be perfectly classified, there is no remaining error to drive the feature learning. If participants are trying to learn about the concept, however, and not simply to classify exemplars, they should learn more about a category than a single, perfectly predictive feature.

As will be seen (in Experiment 3, in particular), we are not questioning the existence or importance of error-driven learning. It is very likely the explanation for blocking and related effects in associative learning (see Kruschke, Kappenman, & Hetrick, 2005). Furthermore, we assume that many of the same associative learning mechanisms are involved in category learning, as computational models have proposed. Our hypothesis is that within a category-learning task, other mechanisms can override error-driven learning and overcome some of its limitations. We withhold more detailed speculation on how this might occur until after presentation of the experiments.

Overview of Experiments

We present three experiments that investigate blocking in category learning. Participants learned to classify novel exemplars into two groups, after which they were tested on their learning of individual features. The exemplars consisted of eight binary features taken from the domain of vehicles. Participants learned either a *defining-dimension* category structure or a *control* category structure. The control group learned categories with a family-resemblance structure in which no single dimension was perfectly predictive of the category. The defining-dimension structure was similar, but crucially, one of the dimensions, the *defining* dimension, was perfectly diagnostic of the category label. Participants who learned the defining-dimension category structure were also given *pretraining*, in which they were trained on the defining dimension prior to being exposed to the whole category exemplars. The pretraining corresponds to the presentation of the light, which perfectly predicted the shock, in the first training phase of Kamin's (1969) study. Participants therefore needed to learn only the defining feature of each category in order to classify all of the exemplars correctly. We removed any participant who did not score perfectly on this feature in the test phase. This manipulation maximized the possibility that participants in the pretraining condition would use the defining feature to classify the exemplars.

This design obviously deviates from the classical conditioning paradigm in a number of respects, including the use of categories rather than shocks (or another unconditioned stimulus) as the outcome to be learned, the presence of multiple features associated with each category (rather than two), and the semantic nature of the stimuli (features of vehicles). However, all of these differences are part of category learning, which involves learning multiple properties of meaningful categories. Therefore, if principles of associative learning are involved in category learning, they should apply in spite of these differences. Furthermore, Experiment 3 showed that blocking effects can be obtained with the same stimuli and procedure.

Prior research on rule use in category learning (e.g., Nosofsky et al., 1994) has found evidence that participants use unidimensional rules in preference to multidimensional rules and that they learn exemplars more quickly when they are able to use unidimensional rules than when they cannot. This suggests that pretrained participants will learn in fewer trials than control participants. The effects of using the defining feature should also be apparent in the tests of individual features after learning. If learning on the nondefining features is blocked, the pretraining group should learn fewer features than the control group does. More interestingly, error-driven learning predicts that participants in the pretraining conditions should learn little about the nondefining features, since after perfect classification has been achieved using the defining feature, there is no error to drive learning on the other features. More specific predictions are made on an experiment-by-experiment basis.

Experiment 1

Participants learned to classify exemplars into two categories of vehicles, A or B. Exemplars were described on eight binary dimensions, four of which were *knowledge* dimensions and four of which were *rote* dimensions (Murphy & Allopenna, 1994). (Experiments 2 and 3 used only rote features.) Knowledge dimensions were related to a common theme. For example, the Category A features *goes on glaciers, heavily insulated,* and *made in Norway* describe a cold climate vehicle, whereas the Category B features *goes in jungles, lightly insulated,* and *made in Africa* together describe a hot climate vehicle. Rote dimensions did not correspond to a known theme. For example, the feature *license plate in front* can occur in either hot or cold climate vehicles without appearing incongruent (see Table 1 for a full list of features).

Participants were randomly assigned to pretraining and control conditions. In the pretraining condition, a single feature was perfectly predictive (necessary and sufficient) of each category. In the *rote pretraining condition*, a rote dimension was defining, and in the *knowledge pretraining condition*, a knowledge dimension was defining. In the learning phase, participants classified exemplars until they were able to classify one block without error. In the subsequent feature-testing phase, they classified individual features without feedback.

Participants in the pretraining condition first went through a prelearning phase, in which they learned to predict each category from a single (defining) feature—that *goes on glaciers* predicted Category A and *goes in jungles* predicted Category B, for example. Participants saw 14 trials on which they had to assign the correct category label and received feedback on their responses, just as in the learning phase. The control group also went through a prelearning phase, but instead of defining features they learned which key to press for each category. For example, they might see the sentence *Press the category A key* and would have to press the

Table 1
Vehicle Features Used to Construct Stimuli

Dimension type and number	Category A	Category B
Knowledge		
1	Used on mountains	Used on safaris
2	Goes on glaciers	Goes in jungles
3	Made in Norway	Made in Africa
4	Heavily insulated	Lightly insulated
Rote		
5	Has air bags	Does not have air bags
6	License plate in front	License plate in back
7	Has cloth seat covers	Has vinyl seat covers
8	Has manual transmission	Has automatic transmission
9	Has CD player	Has cassette player
10	Has rear wheel drive	Has front wheel drive
11	Has a small trunk	Has a large trunk
12	Has two doors	Has four doors

Note. Dimensions 1 to 8 were used in Experiment 1, and Dimensions 5 to 12 were used in Experiments 2 and 3.

appropriate button. After the prelearning phases, all groups classified complete exemplars in the learning phase.

Assuming that participants in the pretraining condition could already classify the exemplars perfectly from the start of the learning phase, error-driven learning would predict that they would perform at chance on the nondefining features because there is no error to drive learning of more features. For example, Gluck and Bower's (1988) connectionist model of category learning would place a high weight on the defining feature in the prelearning phase, but then, because the model could correctly classify every exemplar before the learning phase ever started, there would be no errors to cause the model to include the nondefining features. In contrast, if learning is not entirely error driven, participants in the pretraining conditions would learn something else about the category in the learning phase beyond what they had already learned from prelearning (i.e., beyond the defining dimension).

Increased learning would be visible in two ways. First, participants in the pretraining condition could score better than chance on the nondefining features. Second, participants could extract the theme of hot versus cold climate vehicles. Theme extraction requires learning at least one knowledge dimension because the rote dimensions provide no thematic information. However, a single thematic dimension is not generally sufficient to evoke the theme (see Spalding & Murphy, 1999), and so evidence that learners noticed the theme would reflect learning multiple thematic dimensions. Evidence that participants in the rote pretraining condition extracted the theme therefore indicates that they learned more about the categories than the defining feature.

Blocking is defined as the difference between the experimental and control groups, as just described. However, another interesting measure is how many features of each category participants learn (Hoffman & Murphy, 2006), especially in the pretraining condition. If error-driven learning were the only mechanism involved, then after pretraining, participants would not learn any additional dimensions. Such a strong result is perhaps not expected, but the measure of how many additional dimensions learners acquire indicates the absolute strength of the blocking effect. If people learn four more dimensions, that would suggest a weak blocking effect; if they learn only one additional dimension, that would suggest a stronger blocking effect. To measure the number of dimensions learned, we used a guessing correction, since participants would get half of the features correct simply by chance. Therefore, we subtracted the percent incorrect from the percent correct for each participant to calculate the percentage of dimensions learned. (So, if a participant got only 50% of the features correct on the final test—i.e., chance—subtracting the 50% incorrect would result in an estimate of 0 dimensions learned.) That percentage was multiplied by the number of dimensions to derive the absolute number of dimensions learned.

Method

Participants. Seventy-six New York University students participated for course credit or payment. They were randomly assigned in equal numbers to the control, rote pretraining, and knowledge pretraining conditions. Five other students replaced participants who failed the inclusion criteria, as outlined below.

Stimuli and design. Exemplars were constructed from eight binary dimensions that described Category A and Category B vehicles. Four dimensions were rote and four were knowledge dimensions, based on those from Murphy and Allopenna (1994). Table 1 (in which Dimensions 1 to 8 apply to this experiment) provides a full list. All participants saw 16 exemplars.

The exemplars in the control condition followed the "one-away" design shown in Table 2. The left half of the table refers to Category A, and the right half to Category B. For each category, the rows refer to exemplars and the columns to dimensions. The *1*s and *0*s correspond to the features of the dimensions listed in Table 1. For example, the third row on the left side of the table refers to the third exemplar in Category A. This exemplar consists of seven features that are typical of Category A (*1*s) and one that is typical of Category B (*0*).

For the control condition, no single dimension was completely diagnostic of the category. Therefore, perfect classification performance required knowledge of at least three dimensions, so that every exemplar would have a majority of features from the correct category. This was not the case for the pretraining conditions, in which the defining dimension perfectly predicted category membership. The pretraining conditions therefore had a category structure in which one of the dimensions in Table 2 contained all *1*s for Category A and all *0*s for Category B. In the knowledge pretraining condition this was a knowledge dimension (i.e., one of Dimensions 1 to 4 in Table 1); in the rote pretraining condition it was a rote dimension (one of Dimensions 5 to 8 in Table 1). The precise dimension was rotated across participants.

Procedure. Participants first completed a prelearning phase. Pretraining participants assigned a category label to an individual feature presented on the screen. They received feedback on their responses so that they learned which feature corresponded with which category. These participants saw both features of the defining dimension, but no others. Participants in the control condition went through a similar phase in which they learned to press the key that corresponded to the category label. All participants saw 14 examples (one block) during this phase and then proceeded to the learning phase.

Participants were told that they would be learning about two categories of vehicles and that they would have to learn to classify

Table 2
Category Structure for the Control Condition of Experiments 1 and 2

	Category A									Category B							
Exemplar	D1	D2	D3	D4	D5	D6	D7	D8	Exemplar	D1	D2	D3	D4	D5	D6	D7	D8
1	0	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0
2	1	0	1	1	1	1	1	1	2	0	1	0	0	0	0	0	0
3	1	1	0	1	1	1	1	1	3	0	0	1	0	0	0	0	0
4	1	1	1	0	1	1	1	1	4	0	0	0	1	0	0	0	0
5	1	1	1	1	0	1	1	1	5	0	0	0	0	1	0	0	0
6	1	1	1	1	1	0	1	1	6	0	0	0	0	0	1	0	0
7	1	1	1	1	1	1	0	1	7	0	0	0	0	0	0	1	0
8	1	1	1	1	1	1	1	0	8	0	0	0	0	0	0	0	1

Note. The left side of the table corresponds to the exemplars in Category A, and the right side to Category B. Exemplars are shown as rows in the table, and dimensions as columns. Feature values are shown as 1 (typically Category A) or 0 (typically category B). D = Dimension.

the exemplars by paying attention to the feedback. They were not given information about the theme or defining features.

Exemplars were presented as lists of features, in a different random order on each trial. Participants read the features on a monitor and pressed the key corresponding to one of the categories. They then received feedback indicating whether they were correct, together with the complete exemplar and the correct category. Feedback lasted for 5 s. Learning proceeded in blocks consisting of the presentation of all the exemplars shown in Table 2, in a random order. The learning phase ended when a participant succeeded in classifying all the exemplars of a block correctly, or after 14 blocks.

In the subsequent feature testing phase, features were presented individually on the screen and participants pressed a key to indicate whether the feature was most likely to be from Category A or B. Feedback was not provided. Participants classified each feature twice, for a total of 32 test trials.

Results

Partipants were generally very accurate in the prelearning phase, scoring a mean of 0.97 ($SD = 0.04$) in the control condition and means of 0.92 ($SD = 0.07$) and 0.92 ($SD = 0.06$) in the knowledge and rote pretraining conditions, respectively. Thus, after perhaps one or two errors in the first trials, participants learned the defining feature. (Only those participants who learned their defining feature were included in our analyses; see below.)

In the main learning phase, most participants in the pretraining conditions achieved perfect classification performance after a single block, reaching criterion in 1.29 ($SD = 1.26$) and 1.46 ($SD = 1.67$) blocks for the knowledge and rote pretraining conditions, respectively. The control group required more blocks to learn the categories ($M = 4.00$, $SD = 1.98$), and 2 participants were replaced for failing to reach the learning criterion. (Because there was no qualitative difference between the results of the learning phase with and without the nonlearning participants, we report only the results of those who successfully learned.) The difference between the pretraining and control conditions was highly reliable, $F(2, 69) = 27.46$, $MSE = 2.01$, $\eta^2_p = 0.44$, $p < .001$. This is evidence that participants in the pretraining conditions were using the defining feature and that the pretraining structure facilitated learning. Thus, these are exactly the conditions in which error-driven learning predicts there should be little (if any) additional learning.

We analyzed the single-feature tests separately for the defining features, the nondefining rote features, and the nondefining knowledge features. To maximize the blocking effect (which depends on learning the defining feature), we replaced any participant who did not perform perfectly on the defining feature (as we did in Experiments 2 and 3). There were only 2 such participants in the current experiment. Thus, by design, all participants in the pretraining conditions scored 100% on the defining features, and so those features were not included in any analysis.

Table 3 displays the mean proportion correct for the nondefining rote and knowledge features as a function of the category structure. For the control group, rote and knowledge feature scores were derived from all eight dimensions (see Table 1). However, for the pretraining conditions, one dimension was defining, so scores were derived from only seven dimensions: three rote and four knowledge for the rote pretraining condition and four rote and three knowledge for the knowledge pretraining condition.

Error-driven learning predicts that the performance on nondefining features should be better in the control than in the pretraining group. Surprisingly, the test phase scores gave only marginal evidence that the pretraining manipulation harmed learning on the nondefining features. Those in the control group learned slightly more than those in the pretraining groups: 9% on average, $F(1, 69) = 2.67$, $MSE = 0.06$, $\eta^2_p = 0.07$, $p < .10$. There was no interaction of learning group with feature type, $F(2, 69) = 1.39$, $MSE = 0.03$, $p > .25$. In summary, the blocking effects so strongly predicted by error-driven learning were at best marginal.

One reason for this is that participants in the pretraining conditions learned more nondefining dimensions than predicted. To

Table 3
Proportion Correct (and SDs) of Individual Feature Tests in Experiment 1

Category structure	Rote features	Knowledge features
Control	.75 (.22)	.91 (.09)
Knowledge pretraining	.61 (.19)	.85 (.21)
Rote pretraining	.69 (.24)	.81 (.23)

estimate how many were learned, we first corrected for guessing, as described earlier. Despite not needing the nondefining dimensions to correctly classify the exemplars, pretraining participants learned, on average, 1.0 nondefining rote dimensions and 2.3 nondefining knowledge dimensions in addition to the defining dimension, and accuracy was higher than chance (.50) in all three groups on both rote and knowledge features, ts(23) > 2.8, ps $< .01$. As a particularly stringent test of error-driven learning, we analyzed the feature accuracy of those participants who failed to make an error during the complete exemplar training phase, that is, those who correctly classified all of the exemplars in the first block of the main learning phase. This amounted to 18 participants from the knowledge pretraining condition and 17 from the rote pretraining condition. Performance by these participants was similar to the group as a whole: Accuracy was greater on the knowledge features than the rote features ($M = 0.81$, $SD = 0.23$, vs. $M = 0.64$, $SD = 0.22$), $t(34) = 3.95$, $\eta^2_p = 0.31$, $p < .001$; and accuracy was greater than chance on the rote features and the knowledge features, ts(34) $= 3.75$, ps $< .005$. (When the two pretraining groups were considered individually, differences were always in the same direction but they were not always significant.) Using the guessing correction, we calculated that these error-free participants learned 1.52 features in addition to the defining feature. So, participants still learned category features, even though perfect performance had been achieved before those features were encountered.

There was also evidence that participants had extracted the theme in all three conditions, as measured by the greater accuracy for the knowledge features than the rote features (see Table 3). Participants responded reliably more accurately to the knowledge features than the rote features ($M = 0.86$, $SD = 0.19$, vs. $M = 0.68$, $SD = 0.22$), $F(1, 69) = 36.19$, $MSE = 0.03$, $\eta^2_p = 0.34$, $p < .001$. Reliable simple effects were observed in all three groups, ts(23) > 2.22, η^2_ps > 0.18, $ps < .05$. Particularly important is the knowledge effect in the rote pretraining group, in which the advantage of thematic features reveals that participants must have learned multiple knowledge features in order to evoke the theme and associate it to the correct category, even though a rote feature perfectly predicted category membership.

Greater accuracy on the knowledge features could have occurred for two reasons. First, participants could have extracted the theme during learning, which in turn aided learning of the features related to that theme (Rehder & Murphy, 2003). Second, participants could have been generalizing in the test phase from individual features they had learned. For example, perhaps a participant in the knowledge pretraining condition had learned the defining feature *made in Norway* for Category A vehicles. Then, in the test phase, when presented with *used on mountains*, he or she might have considered it more likely that a vehicle that was made in Norway would also be used on mountains. This explanation, however, predicts that reaction times for nondefining knowledge features would be higher than those for the nondefining rote features, because the latter would have been directly learned and the former only inferred at test.

We analyzed the reaction times of the correct responses to establish whether this was the case. We removed responses longer than five standard deviations above the mean (5.3 s) as outliers (resulting in the removal of 4.0% of the data). Knowledge features were classified more quickly than rote features ($M = 1.48$ s, $SD = 0.44$, vs. $M = 2.00$ s, $SD = 0.58$), $F(1, 69) = 69.78$, $MSE = 135{,}794$, $\eta^2_p = 0.50$, $p < .001$, and there was no interaction with group, $F(2, 69) = 1.10$, $MSE = 135{,}794$, $p > .30$. Furthermore, each group showed reliably shorter reaction times for the knowledge features than the rote features, ts(23) > 4.05, ps $< .001$. Thus, differences in accuracy between the knowledge features and the rote features are unlikely to be due to participants generalizing in the test phase; instead, they likely reflect acquisition of the thematic features during the learning phase.

Discussion

Participants in the pretraining conditions learned the categories more quickly than those in the control condition. Indeed, the majority of participants in the pretraining conditions classified exemplars in the training phase without making a single error. They must have therefore used the defining feature immediately in the learning phase. (We also excluded participants who weren't obviously using the defining feature.)

Because pretraining participants all acquired the defining feature, an error-driven learning account predicts that they should have been impaired in learning the nondefining features relative to the control condition; indeed, they might not have learned any information about the nondefining features. Our results did not support this prediction. There were weak blocking effects—participants were slightly more accurate in the control conditions than the pretraining conditions—but these effects were not statistically robust. This is particularly striking given that participants in the control condition had to learn at least three features to successfully classify all the items, whereas the pretraining condition participants had to learn only one (the defining feature).

In contrast, we observed strong evidence that participants had extracted the theme and that they performed more accurately than chance on the nondefining rote and knowledge features. One might argue that this would be expected from those in the knowledge pretraining conditions because they would be able to generalize from the defining feature to others in the testing phase, even if they had learned only the defining feature. However, we observed response times in the knowledge pretraining conditions that were inconsistent with such a strategy as well as effects of the theme in the rote learning condition, in which generalization from the defining feature would not be possible. Most impressively, these results were also observed in participants who made no errors in learning. Because all (supervised) error-driven models require an error signal to drive learning, these participants provide particularly strong evidence that there is more to learning than eliminating classification errors.

One might wonder whether our results are due to the use of the blocking paradigm in the category-learning task. Perhaps participants find it strange to be trained on one stimulus dimension and then presented with many more dimensions in the main learning trials. Or perhaps they do not completely understand that the same categories are involved in the two phases (even though they have the same names). In an experiment not reported here in detail, we attempted to find blocking without the pretraining phase. Instead, one condition had a defining feature for every category (as in Experiment 1), which was always presented first in a fixed list of features. We assumed that learners would quickly acquire the first, perfectly predictive feature, which could then block learning of the other features. (We again excluded participants who did not learn

this defining dimension.) The stimuli and learning procedure were the same as in Experiment 1. In fact, groups with the defining feature did not learn significantly less than the control group, $F(2, 45) < 1$, and they acquired 1.5 rote dimensions and 3.2 knowledge dimensions, on average, in addition to the defining dimension. Thus, we failed to find blocking under conditions that are more similar to the standard category-learning paradigm, as well as in a paradigm modeled after the conditioning task.

We did not observe reliable blocking effects in our experiments, yet many other researchers have. Why are our experiments different from other blocking experiments? One aspect is that unlike most conditioning experiments, but like real categories, some of our features were related to one another through prior knowledge. Possibly the knowledge features encouraged participants to integrate the stimuli and view the exemplars holistically, preventing the blocking process from taking place. As we discussed in the introduction, there is some evidence that introducing knowledge features facilitates learning of all the features of a concept, not just the knowledge features themselves (Spalding & Murphy, 1999). Of course, this possibility does not save the error-driven learning account, which claims that participants in the pretraining conditions should not be trying to learn about other features, however much knowledge might be activated by the defining feature. Furthermore, because most real-world categories *are* related to some knowledge, this would still represent an important limit on error-driven learning. Nonetheless, to assess the sensitivity of blocking to such variables, we conducted an experiment in which we used only rote features.

Experiment 2

Participants learned the same category structures and learning procedures as in Experiment 1. However, in this experiment, all eight dimensions were rote. There were therefore only two conditions: the control condition and (rote) pretraining condition. If pretraining prevents learning of the nondefining features, then those in the pretraining condition would be expected to perform at chance levels on those features in the test phase. Those in the control group would have to learn some of these features in order to correctly classify the exemplars in the learning phase and so should perform above chance.

Method

Forty-eight New York University students participated for course credit or payment. Five other students replaced those who failed the inclusion criteria (see below). Exemplars consisted of eight rote dimensions, instead of four rote and four knowledge as in previous experiments. The new features are shown in Table 1 (Dimensions 9–12). Because there were now no knowledge features, there were only two conditions: the control group and the pretraining group. The design and procedure were otherwise the same as Experiment 1.

Results

As in Experiment 1, learning was faster in the pretraining condition than in the control condition ($M = 2.25$, $SD = 1.96$, vs. $M = 5.54$, $SD = 3.36$), $t(46) = 4.14$, $\eta^2_p = 0.27$, $p < .001$. Five participants failed to reach criterion in the control condition, whereas all the pretrained participants succeeded. Thirteen out of 24 pretrained participants learned in one block of trials.

Using the defining feature harmed learning the other features relative to the performance of the controls: pretrained participants learned 11% less than the controls ($M = 0.68$, $SD = 0.18$, vs. $M = 0.79$, $SD = 0.12$), $t(46) = 2.47$, $\eta^2_p = 0.12$, $p < .05$. Nonetheless, the pretrained participants scored reliably higher than chance, $t(23) = 4.84$, $p < .001$, as did the 13 participants who classified all of the exemplars correctly in the first block of the main learning phase ($M = 0.69$, $SD = 0.20$), $t(12) = 3.48$, $p < .005$. On average, the pretraining group learned 2.52 features in addition to the defining feature.

Discussion

Experiment 2 yielded stronger evidence of blocking than Experiment 1, with a significant decrease in feature learning after pretraining on a defining feature. Thus, the theme in Experiment 1 might have contributed to the additional feature learning.

Knowledge effects, however, cannot be the whole story. Even in this experiment, in which there was no theme connecting the features, the pretraining group learned additional features, even though the majority had only one block of training and made no errors in it. Furthermore, the control group had over five more blocks of experience with the stimuli on average, because they took longer to reach criterion. Therefore, their higher accuracy cannot be attributed completely to the effect of blocking. In the next experiment, we removed this disparity in number of learning trials and addressed several other reasons why blocking might be reduced in category learning.

Experiment 3

The main goal of Experiments 1 and 2 was to use the classic blocking paradigm to explore whether category learning is entirely error driven. However, the blocking effect itself is of intrinsic interest, and the fact that we found it to be so weak in a category-learning task is surprising. But our experiments did not conform perfectly to the traditional blocking paradigm, and so it is possible that one of the differences between our task and the usual task was responsible for the weakness of our observed blocking effects. Therefore, in Experiment 3 we made a number of changes to the paradigm (guided by reviewers' comments) to bring it into closer conformance with the traditional task. Of course, category learning is not exactly the same as classical conditioning, and the two paradigms will never be perfectly identical. But if the principles of associative learning apply beyond conditioning tasks, we should be able to find them in analogous learning situations with stimuli and procedures appropriate to that task.

One important change was to use perfectly predictive features in both conditions. Previously, the pretraining condition had a perfectly predictive feature, but the control condition did not. As a result, the two groups did not view identical items during their

common learning phase, clouding the comparison of the two.[1] The present experiment, therefore, included a perfectly predictive feature in both conditions. A second difference is that we equated the number of learning trials in the two groups, whereas we previously allowed subjects to learn until reaching criterion. This again made the two groups more comparable, and more similar to the conditioning paradigm.

Experiment 3 also addressed an important theoretical question. We began these experiments with the observation that category learning may evoke different learning strategies than are used in simple classical conditioning or associative learning tasks. Hoffman and Murphy (2006) argued that learners are motivated to acquire many features of a category, whereas such a motivation is typically absent in a conditioning context, in which success at predicting the outcome is the goal. However, Experiments 1 and 2 did not manipulate the learning task but simply studied category learning, pointing out the differences between our results and what would be expected from the simpler paradigms. In Experiment 3, therefore, we made a direct test of our proposal by using formally identical tasks that were presented either as category learning or prediction learning. In one version, subjects were instructed to learn two categories whose features were presented on the screen. In the other version, subjects were instructed to learn to predict when the computer would emit a low or high tone, based on features presented on the screen. In both versions, the same exemplars were presented for the same number of trials, associated in the same way to the two outcomes. The only difference was that the exemplars of one category were associated to the category Mobbles (or Streaths) in one condition and to the presentation of a high (or low) tone in the other condition.

Assuming our explanation of the earlier experiments (and Hoffman & Murphy's, 2006, results) was correct, then we expected to obtain an interaction: Within the prediction condition, we expected there would be a robust blocking effect (people would learn significantly less after pretraining compared to control), but within the category-learning condition, there would be no such effect. Such a result would be strong evidence that the imputed task determines subjects' learning strategies and would explain why we found relatively weak blocking effects in the previous experiments.

Method

Participants. Ninety-six New York University students participated for course credit. They were randomly assigned to one of the four conditions created by crossing the factors of task (category vs. prediction) and pretraining on the defining feature (pretraining vs. control). Five other participants in the pretraining condition failed to learn the defining feature and were replaced (as in Experiments 1 and 2). Data from one participant was removed from the analyses because his proportion correct during the transfer phase was more than three standard deviations below any other in the category task. (Most likely, the participant confused the two categories and responded with the opposite category during the test phase.)

Stimuli, design, and procedure. The present experiment differed from Experiment 2 in the following ways: Participants in the prediction task were instructed to "predict what the computer does"; they predicted whether the computer would emit a low or high tone (220 vs. 880 Hz). In contrast, those in the category task were instructed to "learn about vehicle categories." This was the standard category-learning task. Formally, however, the tasks were identical: People predicted an outcome (tone or category label) from features of vehicles; the relationship between features and outcome was identical across the two tasks. To make the categories more category-like, we used the names *Mobbles* and *Streaths*, rather than *Category A* and *Category B*. All conditions, not just the pretraining condition, had a defining feature. All features were presented in a random order on each trial.

Just as in previous experiments, participants in the pretraining condition were pretrained on the defining feature, and participants in the control condition were pretrained on the association between button and category label (category task) or button and tone (prediction task). In the pretraining prediction condition, participants saw the defining feature and pressed a key with a purple sticker on the keyboard (actually the *z* key) if they thought the computer would produce a low tone, or a key with a yellow sticker (the / key) if they thought it would produce a high tone. In the control prediction condition, participants might see the phrase *high tone* and have to press the button corresponding to that tone. Once they responded, participants would hear that tone for 1 s. After the tone sounded, the corresponding text, *high tone* or *low tone*, appeared with feedback, *CORRECT* or *INCORRECT*, for 5 s. The procedure was identical for the category conditions, but instead of a tone, participants saw the category labels *Mobble* or *Streath*, as in Experiments 1 and 2. Pretraining lasted 14 trials for all participants.

After pretraining, participants were trained for four blocks on the category and prediction tasks, receiving full exemplars rather than single features. However, a preliminary analysis of performance for the first 32 participants indicated the possibility of floor effects in the prediction condition, which would have made any differences between blocking and control undetectable. To minimize this possibility, we increased the number of training blocks from four to six for the remaining 64 participants. Number of training blocks (four vs. six) was therefore included as a between-subjects factor in our analysis. After training, participants were tested for two blocks on individual features, as in Experiments 1 and 2.

Results

Learning phase. We first examined proportion correct throughout learning to see how pretraining reduced the error signal relative to controls. Figure 1 shows average proportion correct as a function of pretraining, task (category and prediction are in the left and right panels, respectively), and experiment block. The figure illustrates that, compared to controls, pretraining participants on the defining feature reduced their error rates throughout training. For both prediction and standard category-learning tasks, the pretrained participants had the highest average proportion correct. However, Figure 1 also shows that the effect of blocking

[1] Actually, such a difference would seem to work in favor of finding a blocking effect, because the controls had to learn multiple features in order to acquire the category, whereas the pretrained participants had to learn just one feature. And note that this difference cannot explain the pretraining group's unnecessary learning of features.

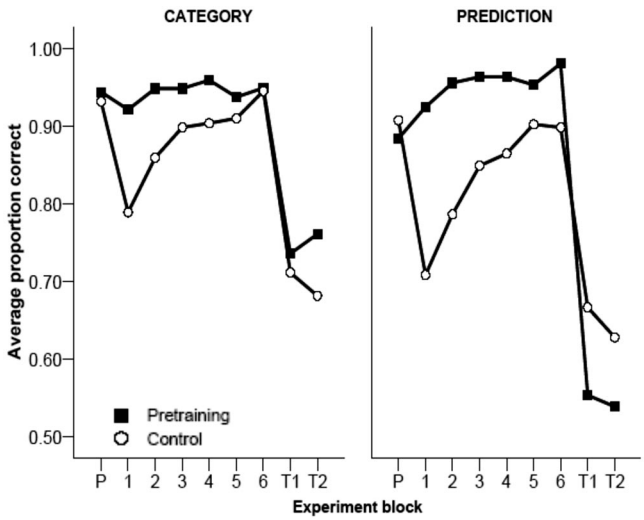

Figure 1. Average proportion correct as a function of task, training block, and pretraining for Experiment 3. Pretraining conditions are indicated by filled squares, and controls by open circles. The blocking effect is revealed in the prediction condition (right panel), in which the pretraining condition is lower than the control condition in the two test blocks. The effect is not present in the category-learning condition (left panel). P = prelearning phase; T1 and T2 = test blocks.

was higher in the prediction task (i.e., the difference between pretraining and control was larger for participants predicting a tone compared to those learning categories).

We tested the reliability of these effects with a 4 × 2 × 2 mixed-design analysis of variance (ANOVA), in which training block was a within-subjects factor and pretraining and task were between-subjects factors. (To include all participants, we used only the first four learning blocks.) We found reliable main effects of training block, $F(3, 273) = 17.76$, $MSE = 0.01$, $\eta^2_p = 0.33$, and pretraining, $F(1, 91) = 55.97$, $MSE = 0.02$, $\eta^2_p = 0.38$, ps $< .01$, confirming that there was an overall improvement in performance and that average proportion correct was indeed higher in the pretraining conditions compared to controls. There was a marginally reliable effect of task, $F(1, 91) = 2.94$, $MSE = 0.02$, $\eta^2_p = 0.03$, $p < .10$, but most importantly, there was also an interaction between task and pretraining, $F(1, 91) = 4.84$, $MSE = 0.02$, $\eta^2_p = 0.05$, $p < .05$, reflecting the stronger blocking effect in the prediction task.

The superior performance in the pretraining conditions depended on training block, $F(3, 273) = 5.93$, $MSE = 0.01$, $\eta^2_p = 0.06$, $p < .01$, reflecting that the average proportion correct in the control conditions improved relative to the pretraining conditions. Whereas the controls began training at chance performance, participants in the pretraining conditions performed near ceiling throughout learning, leaving little room for improvement, and importantly, little error signal to drive additional learning. No other interactions were reliable (Fs < 1).

To better understand the interaction between task and pretraining, we examined simple effects of proportion correct over all training blocks. We found that the interaction was driven by the controls having a higher proportion correct in the category task ($M = .87$, $SD = .10$) than in the prediction task ($M = .82$, $SD = .08$), and this 5% difference was nearly reliable, $t(46) = 1.98$, $\eta^2_p = 0.08$, $p = .054$. However, performance of the pretrained participants in the category ($M = .95$, $SD = .05$) and prediction conditions ($M = .96$, $SD = .05$) were both at ceiling and not reliably different, $t(45) < 1$. Thus, the larger blocking effect for the prediction task compared to the category task derives from the higher performance of control participants who learned categories compared to those who predicted the tones. Note that error-driven learning models have no way of explaining the superior learning performance of the control category group compared to the control prediction group. As far as the models are concerned, the tasks should produce identical performance.

Across tasks, pretraining the participants on a defining feature greatly improved their performance over the course of learning so that the error signal available to drive feature learning in that group was near zero. According to error-driven learning accounts, pretraining the defining feature should have blocked learning about additional features.

Test phase. We next tested for evidence of blocking at test. Figure 1 (left panel, T1 and T2) shows the average proportion correct on the nondefining features from the two test blocks in the category task. Contrary to error-driven learning accounts, there was no evidence of a blocking effect in category learning; the controls did not perform better than those in the pretraining condition. In fact, the effect is in the opposite direction. Moreover, the performance levels indicate that participants learned more than was necessary to do the task.

In contrast to the category-learning task, the prediction task yielded a clear blocking effect. Figure 1 (right panel, T1 and T2) shows the test results for the prediction task, revealing that the controls scored a higher average proportion correct than those in the pretraining group. In contrast, the pretraining group performed at chance levels on the nondefining features. Consistent with a blocking effect, after learning to use the defining feature, participants learned little else.

We examined the reliability of these effects with a three-factor, between-subjects ANOVA (pretraining vs. control, category-learning vs. prediction task, and number of training blocks: 4 vs. 6), averaging over the two test blocks. We found a main effect of task, $F(1, 87) = 16.39$, $MSE = 0.03$, $\eta^2_p = 0.16$, $p < .01$, indicating superior performance in the category task, mitigated by a reliable Pretraining × Task interaction, $F(1, 87) = 5.69$, $MSE = 0.03$, $\eta^2_p = 0.06$, $p < .05$. All interactions and the main effect involving the factor of number of training blocks were not reliable (Fs < 1) except for a marginally reliable interaction between number of training blocks and task, $F(1, 87) = 2.81$, $MSE = 0.03$, $\eta^2_p = 0.03$, $p = .097$, reflecting that learners in the prediction task performed slightly better after the six blocks compared to four.

Planned contrasts revealed that the interaction between task and pretraining resulted from superior performance of the controls compared to the pretrained participants in the prediction task, $F(1, 91) = 4.55$, $MSE = 0.03$, $\eta^2_p = 0.05$, $p < .05$, but no such difference in the category task, $F(1, 91) = 1.20$, $p = .28$. Overall, performance in the control category task was not significantly better than the control prediction task, $F(1, 91) = 1.08$, $p = .30$. Importantly, those in the pretraining prediction condition were the only participants to not perform better than chance on the nondefining features ($M = .55$, $SD = .19$), $t(23) = 1.20$, $p = .24$, indicating the strongest blocking effect across the four experi-

ments. Once learners were able to predict the tone with perfect accuracy with the defining feature, the remaining predictive features were not learned. Contrast this failure to learn any additional features with the pretraining category condition, in which participants learned a large proportion of the nondefining features (M = .75, SD = .13; approximately 3.5 additional features), in spite of their ability to perfectly predict the category on the basis of the pretrained defining feature.

Discussion

The results were important in two respects. First, when the learning conditions were equated more exactly than in Experiments 1 and 2 (both pretraining and control conditions having perfectly predictive features in the category exemplars and the same number of learning blocks), we found no evidence of blocking in the category-learning task. As in Experiments 1 and 2, participants in the pretraining and the control conditions learned more than they needed to in order to classify the exemplars. That conclusion is all the stronger given that blocking did occur with the same exemplars viewed for the same time in the prediction task.

The interaction of task and pretraining condition is strong evidence that learners' construal of the task determines what they learn. Subjects who believed that they were trying to predict the computer's behavior tended to learn the minimum amount possible: They were essentially at chance (55%) in learning the nondefining features once they had learned the perfectly predictive feature. Participants in the corresponding category-learning condition were 75% correct on the same features. Furthermore, there was a blocking effect only in the prediction condition. We suspect that this interaction effect would have been even greater if subjects had been trained until criterion. The prediction task led to much less learning in general, which limited the size of the blocking effect that could be obtained. If the prediction subjects had been forced to learn until criterion, the control subjects would have had to learn more of the features, but the blocking subjects would have generally stopped after a single block of perfect performance (relying on the blocking feature). Thus, although the present experiment does equate the amount of exposure to the features, it probably reduced the blocking effect in the prediction condition relative to a more typical category-learning task.

Finally, an interesting finding was the interaction of task and pretraining condition during the learning trials (as shown in Figure 1). The pretrained participants did well in both tasks, averaging over 90% correct. What is perhaps surprising is that the controls engaging in category-learning performed nearly as well, especially after the first training block. In contrast, controls in the prediction task never caught up to their pretrained counterparts. Clearly, there is something about trying to predict a computer's behavior that leads to less successful learning than trying to learn a category. It is tempting to suggest that the category-learning task is somehow more realistic than predicting what a computer will do. But given that our categories were clearly artificial entities constructed by experimenters, and given that computers' behavior is actually something that our subjects probably try to predict every day, that explanation does not appear very satisfying. Instead, as we discuss below, we suspect that the category-learning task engages attention in a way that prediction learning doesn't.

General Discussion

Most models of category learning employ an error-driven mechanism to explain how people learn categories. Minimizing the difference between actual output and desired output is the driving force behind change in the knowledge structures of such models, and without error, no learning is predicted. Such a learning mechanism is most famously illustrated in the blocking paradigm (e.g., Kamin, 1969), in which one feature is sufficiently predictive of an outcome that learning of other features of the situation is greatly reduced or nonexistent. We argued that such learning is unlikely to be the whole story when real categories are concerned, however, because category use has functions that may require more features than the minimum necessary for classification.

To test our prediction, we adapted the blocking paradigm for a classification task involving family resemblance categories by pretraining participants on a defining feature. Participants in the pretraining conditions were encouraged to use a defining feature that allowed them to perfectly classify the exemplars, whereas control participants either had no such feature (Experiments 1 and 2) or were not pretrained on the feature (Experiment 3). If error-driven learning were solely responsible for feature learning, participants should have learned the minimum necessary to perform the task and no more. Instead, pretrained participants learned more than the minimum number of features necessary for classification of the exemplars, contrary to an error-driven learning account. Indeed, this effect was so strong that we failed to observe a reliable difference in how many features of a category people learned between our pretraining and control conditions in Experiments 1 and 3. We also observed that participants in the pretraining conditions were able to relate the features of the exemplars to each other and extract the theme connecting the exemplars (when one was present), thereby providing further evidence of multiple feature learning and abstraction. Furthermore, all of these effects were present in participants who correctly classified the exemplars without making a single error in the main training phase. Because there was zero error for these participants, this is particularly strong evidence against error-driven learning. These findings are all the more surprising given that we observed strong blocking effects with these materials when participants learned to predict a computer beep rather than a category (Experiment 3).

Our results imply that when people learn categories they are not driven solely by a need to maximize classification accuracy. Instead, they are likely to learn as much information as they can about the exemplars, presumably to maximize performance in other category uses such as inference or induction. This is not to say that classification error plays no part in driving learning; rather, other goals operate *in addition* to minimizing classification error. The following sections elaborate on what these additional goals might be and how they might be instantiated in models of category learning.

Implications for Models of Category Learning

Our goal in this article was not to test individual models of category learning but to test the concept of error-driven category learning in general. Nonetheless, a discussion of how individual models might explain our results is informative because it clarifies why error-driven learning is insufficient to account for learning in

our experiments, and which other strategies our participants could have used.

First, consider how ALCOVE (Kruschke, 1992) might be used to model Experiment 2, say. During the prelearning phase of Experiment 2, ALCOVE would learn that a high attention weight should be placed on the defining dimension and that the association weights should map feature values to the appropriate categories. In the main learning phase, more exemplars were encountered. One could argue that these new exemplars are different from the exemplars encountered in the prelearning phase (i.e., they are represented as separate basis units) or that they are enlarged versions of the same exemplars. In either case, classification accuracy would be perfect in the pretraining condition, because the high weight on the defining dimension would apply to the exemplars, which contain the perfectly predictive features. Because the weights in ALCOVE are changed in proportion to the classification error (weights are derived according to backpropagation principles; Rumelhart, Hinton, & Williams, 1986), no learning would take place during the main learning phase of the experiment. Accuracy in the individual feature test would therefore depend on the weight values at the end of the prelearning phase, namely, high on the defining dimension and zero on the other dimensions. Thus, the model would not learn features other than the defining feature. All models that adjust weights using some variant of backpropagation would make the same predictions.

ALCOVE does not incorporate prior knowledge into its learning, and so it cannot explain how participants extracted the theme in Experiment 1. Models that have been specifically developed to account for prior knowledge effects in category learning, such as BAYWATCH (Heit & Bott, 2000) or KRES (Rehder & Murphy, 2003), fare no better in predicting knowledge effects in the pretraining condition, however. These models assume that knowledge effects arise as a function of the error between the output of the model and the teaching signal. For example, BAYWATCH assumes that known concepts are recruited in learning new categories to the extent that they reduce the error signal. If there is no error, the relationship between prior concepts and new information does not develop. Hence, if the exemplars can be perfectly classified at the end of the prelearning phase, no themes are predicted to develop.

Hypothesis testing models, such as RULEX (Nosofsky, Palmeri, & McKinley, 1994), suggest a different strategy for learning exemplars based on rules. In our experiment, pretrained participants learned exemplars from a structure that can be classified according to a simple rule such as "If the vehicle is made in Norway, then it's an A vehicle." RULEX assumes that people initially start searching for one-dimensional rules that correctly classify the exemplars, and then, if this strategy fails, they augment the rule mechanism with exceptions to the rule or search other one-dimensional and multidimensional rules. The rule search mechanism in RULEX is essentially error driven: A given rule will be perpetuated until it fails to predict the classification of an exemplar during training. This implies that in our experiments, RULEX would only learn about the defining dimension, because there would be no reason to switch from this dimension to explore other features of the exemplars, as the rule is always correct. In its current instantiation then, RULEX would be unable to predict our results.

Most supervised models of category learning can be considered variants of one of the models above. In all such models, relating learning to the goal of reducing classification error means that the models cannot depart from the strategy developed in the prelearning phase and hence cannot learn information about the remaining features of the exemplars. Because the findings from our category-learning experiments are difficult to explain using error-driven, supervised models, it appears that participants might have been using some unsupervised learning mechanism to acquire knowledge of the nondefining features. This suggests that a model with unsupervised learning capabilities, like SUSTAIN (Love, Medin, & Gureckis, 2004), might provide a better explanation of our results.

SUSTAIN learns to categorize exemplars by forming clusters of exemplars that are similar and mapping those clusters onto category labels. New clusters can be formed if feedback indicates that an incorrect cluster assignment has been made (supervised learning) or if a new exemplar is sufficiently different from the existing clusters (unsupervised learning). Adjustable weights connect the input dimensions to the clusters so that dimensions that are useful for categorization can be emphasized over those that are not. Importantly, these weights are not adjusted in proportion to the classification error (unlike ALCOVE, or any other backpropagation model; see Equations 12 and 13 in Love et al., 2004), but in proportion to the distance between the input pattern and the cluster center. Thus, mere exposure to exemplars would lead to learning even after perfect classification performance had been obtained. If SUSTAIN were simulating our category-learning experiments, it would learn about both the defining and the nondefining features because the unsupervised learning component would operate in the main learning phase, regardless of the classification error. SUSTAIN is therefore likely to reproduce the weak blocking effects we observed in our category-learning experiments. Of course, displaying a weak blocking effect when modeling category-learning experiments suggests that it is likely to predict weak blocking effects in other blocking paradigms, such as that used by Kruschke and Blair (2000) or our computer prediction task in Experiment 3.

To summarize, we know of no model of category learning that, in its published form, could reproduce the behavior of participants in our experiments. To do so would require two learning mechanisms in addition to the error-driven learning instantiated in most models. First, a model must include some form of learning that is independent of the discrepancy between the output of the model and the teaching signal. This could be an unsupervised learning rule, like SUSTAIN's. Second, the model must be able to use different learning mechanisms for category learning versus simple prediction learning. Third, the model must be able to incorporate prior knowledge into learning. Any model that seeks to explain why participants learn more than necessary in a blocking paradigm must also be able to explain why this effect was strongest when features were thematically related.

Behavioral Versus Internal Error

In discussing error-driven learning, we have focused on what might be called *behavioral error*—simply, whether the learner gets the category correct or incorrect on a trial. However, from the perspective of many models, there is also *internal error*—the

difference in activation between a given node and its desired activation. For example, imagine that in response to an exemplar, a model's nodes for Categories A and B achieved activation values of .75 and .25, respectively. The model might then be interpreted as selecting Category A and therefore as not making an error on that trial. However, if the correct choice was Category A, the target output might be 1 for that node and 0 for the Category B node. Therefore, even though the model could be responding perfectly correctly, it could still have internal error that drives learning of other features. Could this type of error be explaining the learning in our experiments?

We cannot rule out this possibility conclusively, as the answer depends on just how each model arrives at a response. However, most models do not behave like the example above. If Output Nodes A and B have activations of .75 and .25 (on a 0–1 scale), then most models would not in fact always choose Category A as the correct answer. Instead, since Medin and Schaffer (1978), models have tended to make decisions stochastically, often following something like Luce's Choice Axiom (Nosofsky, 1984). That is, the model might choose Category A 75% of the time for such a stimulus, rather than always choosing Category A because it is the most likely. Thus, such models would not have zero behavioral error along with nonzero internal error—they would continue to make behavioral errors when the activation was less than the target amount.

More recently, models have adopted a specific parameter, called *gamma*, that specifically varies this aspect of the decision rule (Ashby & Maddox, 1993). If gamma is high, then the model behaves more deterministically (i.e., the most active node determines the response); if gamma is low, then the model does probability matching (i.e., choosing A 75% of the time in the above example). It is only for high values of gamma that this difference between behavioral and internal error can arise. Under these circumstances, a model using error-driven learning might have near-perfect behavioral accuracy but still be learning because of internal error that is not apparent in the behavioral decision. Whether specific models can in fact set their parameters so that they can capture our phenomena has yet to be seen. But most models do not have deterministic responding (high gammas) in published experiments. High gammas are usually cited when there is an unusual finding to be explained. Furthermore, we think that such internal error is unlikely in the blocking paradigm. It is unclear why there would be significant internal error after learning only two features in the first phase, long after classification became perfect.

It also possible that different types of internal error signal might lead to greater degrees of learning on the individual features. One possibility is that at least some part of the overall error would be determined by the error between each feature's category prediction and the category label (whether Feature 1 is predictive of the category label, whether Feature 2 is predictive, whether Feature 3 is predictive, etc.). If models were constructed that learned by maximizing individual feature predictability, rather than maximizing classification performance, the internal error signal would continue to be nonzero even after the exemplar could be correctly classified. There are two ways in which such a feature-to-outcome error term might be instantiated. The first is to assume that the model would be driven solely by the need to minimize feature-to-outcome errors. This approach could lead to a model that was limited in the range of category structures that it could learn, however, because optimizing feature-to-outcome performance might come at the expense of classification performance (e.g., the global minimum for feature-to-outcome error might lead to misclassifying atypical exemplars). In effect, such a model would not be a model of category learning, but one of feature learning.

A more plausible alternative would be a model that combined a feature-to-outcome error term with a classification error term. Indeed, Kruschke (2001) presents a mixture-of-experts model based on EXIT (Kruschke & Johansen, 1999) that incorporates such an error term. The model assumes that each expert learns to associate one cue (or feature) with the outcome, and the different experts are weighted and combined to produce a classification judgment. Each expert minimizes the error between its own prediction and the true classification, whereas the experts are combined to minimize the overall classification error. The individual experts might therefore continue to optimize their predictions long after perfect classification has been achieved by the system as a whole. Although promising (Kruschke, 2001, demonstrates that such a model simulates blocking behavior), this particular instantiation of a mixture-of-experts model assumes that learning within each expert takes place in proportion to the extent that it is used in the overall classification decision (see Equation 17, Kruschke, 2001, p. 834). Thus, if one expert (feature) is highly successful in predicting the classification outcome, as is the case in the pretraining conditions of the present experiments, there is very little drive for the other experts to learn the associations. Lifting the restriction on module learning might allow the model to learn additional features, but this might result in an error term that was entirely determined by feature learning, rather than classification learning, encountering the problems involved in learning atypical exemplars described above. (An extension of the model might be to parameterize the degree to which learning is feature determined or category determined, as in Love & Jones, 2006, e.g., but without a proposed learning mechanism for determining the parameter values, the parameterization adds little to the qualitative explanation of our effect.)

We conclude that it may be theoretically possible for models to account for the significant amount of learning of category features that we found through internal error signals combined with deterministic responding, or perhaps through alternative error terms based on feature-to-outcome performance. However, each model must find the parameters that lead to very fast learning of the defining feature, and low or no performance errors on the subsequent category-learning blocks, but enough internal error to drive the learning of a number of features of each category even with only one block of exposure. Finally, the crux of our argument is that participants are learning more than the minimum necessary to classify exemplars. Invoking the use of an internal error term, as distinct from the behavioral error term, acknowledges that there is more to category learning than classification.

Why Learning Categories Is Different From Learning to Predict Outcomes

We experienced difficulty obtaining blocking effects in category learning, yet blocking is a highly replicated and prevalent phenomenon in associative learning. The results of Experiment 3 suggest that one of the reasons for this difference is that other researchers have used outcome prediction tasks, whereas our par-

ticipants learned categories. Our experiments demonstrated two significant findings concerning the difference between learning categories and learning to predict outcomes. First, when people learn categories they learn more features than are necessary to classify the exemplars, whereas they learn these extra features to a lesser extent (or not at all) when learning to predict outcomes. Second, the effects of a blocking manipulation are greater when people learn to predict outcomes than when they learn categories. These two effects might be related, as we discuss below, but they are nonetheless behaviorally dissociable.

One of the differences between learning a category and predicting an outcome is that categories have a hierarchical structure, and therefore afford inferences, whereas outcomes do not. For example, in the category condition of our task, participants learned to classify individual vehicles as a Mobble or a Streath. The utility of knowing the appropriate classification is that if a future vehicle were identified as Mobble, inferences could be made about the likely features of that vehicle, even if they were not known at the time of classification. Learning more features is beneficial because it increases the inference possibilities (as we discussed in the introduction). The situation is different when predicting an outcome, however. For example, when participants learned to predict the behavior of the computer in our prediction task, they were not classifying the input as a high tone or a low tone; they were predicting whether a high tone would occur given a particular list of vehicle features. It is not the case that a list of vehicles *is a* tone, so the task did not involve hierarchical classification. Similar prediction tasks have been used in many other studies on blocking; for example, Chapman and Robbins (1990) used a stock market prediction paradigm, Kruschke and Blair (2000) used symptoms that predict diseases, and Shanks (1985) used different weapons that cause tanks to explode. Without the act of classification, people may have seen little incentive in learning more features.

A further possibility is that when people learn about categories, they try to integrate the new information with prior networks of knowledge. Indeed, a difference between the prediction task and the category task is that the tones used as category labels did not relate in any way to known categories or knowledge in general. For example, because we informed participants that they were learning about vehicle categories in Experiment 3, participants were likely to have searched for prior vehicle concepts to help them learn, whereas participants in the predictive condition may not have had the same incentive to integrate the new information, or they could have had less success if they tried (they might not have known about other instances in which vehicle features cause computers to beep). Research into the effects of prior knowledge on category learning has consistently demonstrated that features that are easily integrated with prior concepts are easier to learn (e.g., Heit & Bott, 2000; Murphy & Allopenna, 1994; Rehder & Murphy, 2003). Thus, a potential explanation for why participants in our studies learned more in the category condition than the prediction condition is that the former found the features easier to integrate with their prior knowledge. This might also explain why we observed (small) blocking effects in Experiment 2, in which the categories were knowledge poor, but not in Experiment 1, in which the categories were knowledge rich.

In addition to learning more overall, participants learning categories were less affected by the pretraining manipulation than those learning to predict tones. One interpretation of this finding is that people are less content to accept that a category can be defined on a single dimension than they are to accept that outcomes can be successfully predicted by single dimension. They therefore try to learn more dimensions of a category than an arbitrary outcome. This could be for several reasons. First, as we suggested above (and as suggested in Hoffman & Murphy, 2006), if people understand that inference making is a crucial function of categories, they will be driven to learn as much as possible to maximize the usefulness of the category knowledge. Rather than reducing their attention after learning the defining feature, they may continue to attend to the presented features and learn their relationship to the two categories. In contrast, they may be content to believe that outcomes like computer beeps have a single underlying cause and that once successful prediction has been achieved, there is no benefit to learning more.

Second, category-learning and prediction tasks might differ in how the features of the exemplars are perceived to interact. In an outcome prediction task, features might be perceived as mutually exclusive if they are thought of as potential causes of an effect. Waldman and Holyoak (1992) demonstrated that when a set of features were perceived as causes, more blocking occurred than when they were perceived as effects. In causal learning tasks, learning that one object causes an outcome results in discounting of other possible causes associated with the outcome (Sobel, Tenenbaum, & Gopnik, 2004). Similarly, Williams, Sagness, and McPhee (1994) showed that instructions describing features as mutually exclusive resulted in more blocking than neutral instructions. Because exemplars of coherent categories have features that generally reinforce each other rather than being mutually exclusive (Murphy & Medin, 1985), category learning is likely to be less susceptible to blocking than outcome prediction in general.

We have made several suggestions for why learning about categories might result in different knowledge representations than learning to predict outcomes. At least some of the suggestions are perfectly compatible with an error-driven account of learning (e.g., mutually exclusive dimensions could be represented by inhibiting links between dimensions), but our results also demonstrate that the factors affecting performance operate in the absence of classification error. After all, pretrained participants learned more in the category-learning condition than they did in the prediction condition, even when both groups could classify all of the exemplars from the outset.

Relation to Prior Results

Our finding of reduced blocking and increased learning of unnecessary features in category learning appears related to two well-known phenomena in the category-learning literature—one of them seemingly similar and the other possibly conflicting. We discuss the relation of our results to these prior findings here.

The first finding is that of Allen and Brooks (1991; see also Erickson & Kruschke, 1998), who also trained people on a categorization rule and then showed that they learned more about the category than the rule. In particular, the categories were defined by three probabilistic features, but learners also attended to other features that were not predictive of categorization. How does our demonstration differ from theirs?

One important difference between the studies is that Allen and Brooks (1991, Experiment 1) used a procedure specifically de-

signed to lead to strong memory of the individual exemplars. There were only eight training items, which contained colored background scenes unique to pairs of items. Rather than the standard learning trial used here (presenting the item, followed by feedback), Allen and Brooks engaged learners in significant training to ensure their memories of the specific exemplars. The features of each exemplar were presented across three slides, and learners were required to recall the information from previous slides on Slides 2 and 3. This was done "to individuate the animals and to prevent them from being processed solely as instances of the rules" (p. 6). No such training was done in our experiments. Thus, the strong exemplar memory Allen and Brooks documented cannot be generalized to the stimuli and procedure used in most classification learning experiments.

Furthermore, our studies were designed to explore what is learned in categorization, whereas Allen and Brooks (1991) focused on generalization. Indeed, in their critical training condition, learners were simply told the classification rule prior to learning. Their test compared novel exemplars in which the rule feature contradicted the exemplar information to exemplars in which there was no contradiction. In contrast, we collected responses to individual features because we wished to test what information participants had learned, and not what they chose to use in conflict trials. Allen and Brooks did not collect this information. Finally, during their test, the colored backgrounds were presented prior to the rest of the picture, to maximize its possible effect. In contrast, we did not give any special emphasis to one kind of feature over the others. In short, there are a number of important differences between these two paradigms, and the results from one cannot be easily extended to the other. This is not surprising, because the goals of the two experiments were quite different. In fact, it is interesting to note that Allen and Brooks (Experiment 2) failed to find exemplar effects with verbal lists of stimuli, ostensibly because participants were unable to form coherent, memorable units. Thus, our finding of unnecessary learning of verbal features is clearly not redundant with their demonstration. (The difference likely lies in the category structure and test procedures, as described above.)

The second related literature concerns that of *category use*, and in particular the comparison of classification learning and inference learning mentioned in the introduction. A number of studies have concluded that participants learn the minimal amount necessary to ensure classification when they are trained via classification learning, as in the present experiments. In contrast, learning a category by answering inference questions leads to better learning of the category's properties and their relations (see Chin-Parker & Ross, 2002; Yamauchi & Markman, 1998, 2000). Our finding, that people learn more features than necessary to classify correctly, does seem to conflict with these conclusions to some degree. However, it is possible that the two kinds of experiments are simply making different comparisons. We have compared classification learning to prediction learning and found that people acquire more category features in classification learning, even when it is not necessary to do so. This suggests that classification learning does not lead to *minimal* learning of a category's features, but that does not mean that other forms of learning would not be better. It is still possible that inference learning would have led to acquisition of still more properties, as past experiments have found. Indeed, participants learned more features when they were related to prior knowledge (Experiment 1 vs. Experiment 2), again reflecting the fact that participants were not learning the minimal amount necessary. How much will be learned about a category is a complex function of the category structure, acquisition task, relation of the category to prior concepts and knowledge, and the learner's strategy. Although classification learning probably leads to less about the category being learned than some other tasks, when the other variables are propitious considerable learning can still take place with this task.

Conclusion

Learning to classify exemplars through supervised learning is generally assumed to be an error-driven process. People learn what is necessary to classify the exemplars, and they stop learning when they can correctly classify each exemplar. After all, why should they learn any more? In the experiments presented in this article we tested this assumption and found, to the contrary, that a significant amount of learning takes place beyond the point at which classification error is zero. This extra learning means that blocking effects are reduced when people learn categories compared to when they learn to predict other outcomes. To account for our results, current models of supervised category learning need to change their emphasis on classification error so that other learning goals are also included. The challenge for these altered models would be to reproduce the robust blocking effects found in prior studies together with the weak blocking effects and overlearning observed in ours.

References

Ahn, W. (1998). Why are different features central for natural kinds and artifacts? The role of causal status in determining feature centrality. *Cognition, 69,* 135–178.

Allen, S. W., & Brooks, L. R. (1991). Specializing the operation of an explicit rule. *Journal of Experimental Psychology: General, 120,* 3–19.

Anderson, A. L., Ross, B. H., & Chin-Parker, S. (2002). A further investigation of category learning by inference. *Memory & Cognition, 30,* 119–128.

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

Ashby, F. G., & Maddox, W. T. (1993). Relations between prototype, exemplar, and decision bound models of categorization. *Journal of Mathematical Psychology, 37,* 372–400.

Bott, L., & Murphy, G. L. (in press). Subtyping as a knowledge preservation strategy in category learning. *Memory & Cognition.*

Chapman, G. B., & Robbins, S. J. (1990). Cue interaction in human contingency judgment. *Memory & Cognition, 18,* 537–545.

Chin-Parker, S., & Ross, B. H. (2002). The effect of category learning on sensitivity to within-category correlations. *Memory & Cognition, 30,* 353–362.

Dickinson, A., Shanks, D. R., & Evenden, J. L. (1984). Judgement of act-outcome contingency: The role of selective attribution. *Quarterly Journal of Experimental Psychology, 36A,* 29–50.

Erickson, M. A., & Kruschke, J. K. (1998). Rules and exemplars in category learning. *Journal of Experimental Psychology: General, 127,* 107–140.

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

Heit, E., & Bott, L. (2000). Knowledge selection in category learning. In

D. L. Medin (Ed.), *Psychology of learning and motivation* (pp. 163–199). San Diego: Academic Press.

Hoffman, A., & Murphy, G. L. (2006). Category complexity and feature knowledge: When more features are learned as easily as fewer. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32,* 301–315.

Kamin, L. J. (1969). Predictability, surprise, attention, and conditioning. In R. M. Church & B. A. Campbell (Eds.), *Punishment and aversive behavior* (pp. 279–296). New York: Appleton-Century Crofts.

Kaplan, A. S., & Murphy, G. L. (2000). Category learning with minimal prior knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 829–846.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review, 99,* 22–44.

Kruschke, J. K. (1993). Three principles for models of category learning. In G. V. Nakamura, R. Taraban, & D. L. Medin (Eds.), *Categorization by humans and machines: The psychology of learning and motivation* (Vol. 29, pp. 57–90). San Diego: Academic Press.

Kruschke, J. K. (2001). Toward a unified model of attention in associative learning. *Journal of Mathematical Psychology, 45,* 812–863.

Kruschke, J. K., & Blair, N. J. (2000). Blocking and backward blocking involve learned inattention. *Psychonomic Bulletin & Review, 7,* 636–645.

Kruschke, J. K., & Johansen, M. K. (1999). A model of probabilistic category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25,* 1083–1119.

Kruschke, J. K., Kappenman, E. S., & Hetrick, W. P. (2005). Eye gaze and individual differences consistent with learned attention in associative blocking and highlighting. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31,* 830–845.

Lassaline, M. E., & Murphy, G. L. (1996). Induction and category coherence. *Psychonomic Bulletin & Review, 3,* 95–99.

Love, B. C. (2002). Comparing supervised and unsupervised category learning. *Psychonomic Bulletin & Review, 9,* 829–835.

Love, B. C., & Jones, M. (2006). The emergence of multiple learning systems. In *Proceedings of the 28th annual meeting of the Cognitive Science Society* (pp. 507–512). Mahwah, NJ: Erlbaum.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review, 111,* 309–332.

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

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

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

Murphy, G. L. (2005). The study of concepts inside and outside the lab: Medin vs. Medin. In W. Ahn, R. L. Goldstone, B. C. Love, A. B. Markman, & P. Wolff (Eds.), *Categorization inside and outside the lab: Essays in honor of Douglas Medin* (pp. 179–195). Washington, DC: American Psychological Association.

Murphy, G. L., & Allopenna, P. D. (1994). The locus of knowledge effects in concept learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 904–919.

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

Murphy, G. L., & Wisniewski, E. J. (1989). Feature correlations in conceptual representations. In G. Tiberghien (Ed.), *Advances in cognitive science: Vol. 2. Theory and applications* (pp. 23–45). Chichester, England: Ellis Horwood.

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

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review, 101,* 53–79.

Pearce, J. M., & Bouton, M. E. (2001). Theories of associative learning in animals. *Annual Review of Psychology, 52,* 111–139.

Rehder, B. (2007). Essentialism as a generative theory of classification. In A. Gopnik & L. Schultz (Eds.), *Causal learning: Psychology, philosophy, and computation* (pp. 190–207). Oxford, England: Oxford University Press.

Rehder, B., & Murphy, G. L. (2003). A knowledge-resonance (KRES) model of knowledge-based category learning. *Psychonomic Bulletin & Review, 10,* 759–784.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning: II. Current research and theory* (pp. 64–99). New York: Appleton-Century-Crofts.

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* (pp. 111–144). New York: Academic Press.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition: Vol. 1. Foundations.* Cambridge, MA: MIT Press.

Schyns, P. G., & Rodet, L. (1997). Categorization creates functional features. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23,* 681–696.

Shanks, D. R. (1985). Forward and backward blocking in human contingency judgement. *Quarterly Journal of Experimental Psychology, 37B,* 1–21.

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs: General and Applied, 75*(13, Whole No. 517).

Smith, E. E., & Medin, D. L. (1981). *Categories and concepts.* Cambridge, MA: Harvard University Press.

Sobel, D. M., Tenenbaum, J. B., & Gopnik, A. (2004). Children's causal inferences from indirect evidence: Backwards blocking and Bayesian reasoning in preschoolers. *Cognitive Science, 28,* 303–333.

Spalding, T. L., & Murphy, G. L. (1996). Effects of background knowledge on category construction. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22,* 525–538.

Spalding, T. L., & Murphy, G. L. (1999). What is learned in knowledge-related categories? Evidence from typicality and feature frequency judgments. *Memory & Cognition, 27,* 856–867.

Waldmann, M. R., & Holyoak, K. J. (1992). Predictive and diagnostic learning within causal models: Asymmetries in cue competition. *Journal of Experimental Psychology: General, 121,* 222–236.

Williams, D. A., Sagness, K. E., & McPhee, J. E. (1994). Configural and elemental strategies in predictive learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 694–709.

Yamauchi, T., & Markman, A. B. (1998). Category learning by inference and categorization. *Journal of Memory and Language, 39,* 124–148.

Yamauchi, T., & Markman, A. B. (2000). Inference using categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 776–795.

Received May 26, 2006
Revision received April 10, 2007
Accepted April 17, 2007 ■