Prior knowledge and exemplar frequency

HARLAN D. HARRIS, GREGORY L. MURPHY, AND BOB REHDER
New York University, New York, New York

New concepts can be learned by statistical associations, as well as by relevant existing knowledge. We examined the interaction of these two processes by manipulating exemplar frequency and thematic knowledge and considering their interaction through computational modeling. Exemplar frequency affects category learning, with high-frequency items learned more quickly than low-frequency items, and prior knowledge usually speeds category learning. In two experiments in which both of these factors were manipulated, we found that the effects of frequency are greatly reduced when stimulus features are linked by thematic prior knowledge and that frequency effects on single stimulus features can actually be reversed by knowledge. We account for these results with the knowledge resonance model of category learning (Rehder & Murphy, 2003) and conclude that prior knowledge may change representations so that empirical effects, such as those caused by frequency manipulations, are modulated.

Frequency has long been known to be an important property of category structure. Rosch and Mervis (1975) argued that the frequency of properties in a category determines how typical category members are. Those that have properties frequently found in the category are more typical than those possessing less frequent properties, and category members possessing properties frequently found in other categories are less typical than those possessing less frequent properties. Although the frequency or familiarity of an object does not itself seem very strongly related to its typicality in natural concepts (Barsalou, 1985; Mervis, Catlin, & Rosch, 1976; Novick, 2003), when item frequency has been experimentally manipulated independently of other variables (such as similarity to category prototypes), it does influence category structure. For example, Nosofsky (1988) showed that repeating one item five times in each block of category learning made the item not only easier to learn, but also more typical after learning. Furthermore, the effect spread beyond the frequent item itself, in that similar items in the same category also were rated as more typical.

Theories of concepts can explain such frequency effects easily (Barsalou, Huttenlocher, & Lamberts, 1998). If exemplar theories assume that each presentation of a stimulus is a stored instance, frequent exemplars will have more stored instances, increasing the typicality of items similar to them. Likewise, if prototype theories assume that the category prototype is based on generalizing from instances, the more an item is repeated, the more influence it will have on that generalization. That is, frequent items will pull the category prototype in their direction. Thus, the effect of frequency seems to be a straightforward example of how category structure influences learning and use of concepts.

One might expect basic variables, such as frequency, to have consistent effects across materials and tasks. How-ever, there are a number of examples in which the effects of category structure are altered when concepts make contact with other knowledge. For example, the standard learning advantage of conjunctive (*and*) over disjunctive (*or*) concepts can be overruled when the disjunctive concept is related to prior knowledge (Pazzani, 1991). Wattenmaker, Dewey, Murphy, and Medin (1986) examined the effects of prior knowledge on learning linearly separable and nonlinearly separable categories. Linear separability is a structural variable that refers to whether correct categorizations can be made by independently weighting the category's properties. Wattenmaker et al. showed that linearly separable categories could be made easier or harder to learn than nonlinearly separable categories by varying the categories' content (see also Murphy & Kaplan, 2000). They argued that some conceptual domains encourage summing of evidence, suitable for linearly separable categories, and that other domains encourage configural processing, suitable for nonlinearly separable categories. These content effects are one example of how people's prior knowledge about a category can affect the processing they perform during learning and, thereby, alter the influence of structural variables.

In the present research, we examined whether frequency effects are similarly sensitive to the content of the category being learned. Because frequency is such a basic variable, influencing cognitive processes from learning to lexical access, it is possible that its effects will not be so easily modified by the content of a category. We were particularly interested in this structural variable because it allowed us to investigate possible interactions of structural or formal aspects of a category with the more slippery variable of prior knowledge.

We also used this problem of the interaction between knowledge and exemplar frequency to test a model of cat-

egory learning that attempts to incorporate both structure and knowledge, the knowledge resonance (KRES) model (Rehder & Murphy, 2003). Unlike most other models of category learning, KRES allows knowledge, represented by links among features and prior concept nodes, to influence the learning process. We carried out this modeling in order to try to provide an account of the effects of frequency and knowledge in this task and, more generally, to understand better how prior knowledge affects representations during learning. This work also continued our ongoing validation of the model's general approach.

Earlier KRES modeling work (Harris & Rehder, 2006) compared two model variants on linearly and nonlinearly separable category-learning tasks. One model represented prior knowledge by specific nodes that represented already-known categories, which could be associated to the to-be-learned categories. This model could base category responses on the similarity of stimuli to prior concepts. The other variant allowed prior knowledge to influence the learning of associations only by modifying representations of the stimuli. In this second model, knowledge could not directly and independently affect categorization but, instead, had to affect responses by modulating the normal category learning and categorization system. The first variant fit Wattenmaker et al.'s (1986) data better than the second did. However, this may be because Wattenmaker et al. used categories that corresponded to known concepts (e.g., the personality trait of honesty). In the present study, we used what we call *thematic feature relationships* (Murphy & Allopenna, 1994), in which all the knowledge-related features are consistent with a schema or theme but no known category actually exists. Since people often learn new categories that do not correspond to already-known ones, it is important to study and attempt to model this form of knowledge and its influence on learning.

The present experiments, exploring the effects of prior knowledge on category learning with varying exemplar frequency, are a step toward understanding the circumstances under which prior knowledge can affect concept learning and, when combined with computational modeling, will elucidate the mechanism underlying these effects. For expository purposes, we will postpone detailed description of our modeling efforts until after the experiments have been described, so that we may then discuss

the relationships among the data, the model, and the theory in detail.

EXPERIMENT 1

In Experiment 1, subjects saw descriptions of buildings and learned to classify the buildings into two categories. For half of the subjects, most features of the buildings could be linked by themes like *aerial buildings* and *underwater buildings*, whereas for the other half of the subjects, the features were unrelated to each other. (Note that aerial and underwater buildings are not familiar concepts for most people, as was confirmed by Murphy & Allopenna, 1994, and Spalding & Murphy, 1999.) Table 1 shows the stimulus features. To manipulate frequency, one item of each category was presented six times more often than the other items. Once during learning and twice after learning, the subjects performed test trials in which they classified or rated several types of stimuli (trained items, novel prototype items, and individual features), and the effects of knowledge and stimulus frequency on their responses were examined.

On the basis of the work of Nosofsky (1988) and Barsalou et al. (1998), we expected to find frequency effects when prior knowledge was absent. That is, both the frequent exemplar and its features would be more likely to be categorized into the appropriate category than would less frequent exemplars and their features. The category structure we tested is presented in Table 2. This structure follows a standard one-away design in which 11111 is the prototype of Category A, 00000 is the prototype of Category B, and each category member contains one exception feature (a feature characteristic of the other category). However, not all the items in Table 2 were presented an equal number of times. Specifically, 11110 was presented six times more often than the other Category A members, and 00001 was presented six times more often than the other Category B members. Note that this manipulation of exemplar token frequency changes which features are, in fact, associated with which categories. The 0 feature appears more frequently on the fifth dimension in Category A members, whereas the 1 feature on that dimension appears more frequently in Category B members.

To understand this manipulation, imagine that you had a black squirrel living in your back yard. Frequent expo-

Table 1
Feature Pairs Used in the Experiments

Related Features	
divers live there	astronauts live there
get there by submarine	get there by airplane
deep-sea research is conducted there	atmospheric research is conducted there
has thick, heavy walls	has thin, light walls
fish are kept there as pets	birds are kept there as pets

Unrelated Features	
has a large kitchen	has a small kitchen
has area rugs	has wall-to-wall carpeting
has modern furniture	has colonial-style furniture
has a patio	has a porch
has rectangular doorways	has round doorways

Table 2
Abstract Category Structure for Experiment 1

Item	Features	Frequency	Item	Features	Frequency
		Training Items			
A1	1 1 1 1 0	6	B1	0 0 0 0 1	6
A2	1 1 1 0 1	1	B2	0 0 0 1 0	1
A3	1 1 0 1 1	1	B3	0 0 1 0 0	1
A4	1 0 1 1 1	1	B4	0 1 0 0 0	1
A5	0 1 1 1 1	1	B5	1 0 0 0 0	1
		Additional Test Items			
A0	1 1 1 1 1		B0	0 0 0 0 0	
AF1	– – – – 1		BF1	– – – – 0	
AF2	– – – 1 –		BF2	– – – 0 –	
AF3	– – 1 – –		BF3	– – 0 – –	
AF4	– 1 – – –		BF4	– 0 – – –	
AF5	1 – – – –		BF5	0 – – – –	

Note—Item frequency is provided for training items. For test items, A0 and B0 were novel prototype items, whereas AF* and BF* were single-feature items, with AF1 and BF1 being exception features (assuming that A1/B1 were the high-frequency items).

sure to this squirrel, assuming that you do not recognize that it is just a single individual, not only would increase your knowledge of the normal shape, size, and behavior of squirrels but also would incorrectly increase the association between black fur and squirrels (given that the vast majority of squirrels are not black). Thus, high frequency (HF) of an individual exemplar may have a negative effect on learning some of a category's features—those that are idiosyncratic to it. We expected that such HF *exception* features (those presented many times in the "wrong" category) should have low classification accuracy, as compared with the other features. We also expected analogous effects in classification and typicality ratings of the test items.

What should be expected when categories are related to prior knowledge? It is possible that the effects of frequency will be moderated in this condition. In an experiment in which feature frequency, rather than item frequency, was manipulated, Murphy and Allopenna (1994, Experiment 2; see also Spalding & Murphy, 1999, Experiment 3) found much smaller effects of frequency on classification and typicality when prior knowledge was relevant. (They used a somewhat unusual category structure without crossover features that did not allow the analysis we will describe below.) Here, we predicted similar effects for several different reasons. Because learning is faster when prior knowledge is present, empirical manipulations may have fewer opportunities to change what is learned. Furthermore, prior knowledge might tend to counteract some effects of frequency. In particular, the atypical feature of the frequent exemplar might be less influenced by frequency, since it is thematically inconsistent with the rest of the category. For example, perhaps when learning about underwater buildings (although not labeled as such), people might encounter a frequent example that had the exception feature *astronauts live there* (along with four typical features). Although frequency would associate this feature to its incorrect category, people may tend to ignore or downplay the feature, because it does not fit with the category theme, thereby weakening the frequency manipulation (see Heit, 1994). Further analysis of the responses to learned

items and novel stimuli may also address more subtle effects of prior knowledge.

Method

Subjects. Forty members of the New York University community received course credit for their participation. Nineteen subjects were assigned to the knowledge condition, and 21 to the no-knowledge condition.

Stimuli. Each subject saw training examples consisting of the written features in Table 1. The features were based on the integrated (knowledge) and nonintegrated (no-knowledge) feature sets of Murphy and Kaplan (2000), but since we needed additional stimuli, we generated a set of potential additional knowledge-related and knowledge-unrelated dimensions and normed them. Fourteen additional subjects were given lists of features (two values for each dimension, as in Table 1) and were asked how likely it was that each feature would be present in buildings that were either *underwater* or *in the air*. For each pair of items, the likelihood of being in the two types of buildings was calculated. Items that had similar likelihood ratings for underwater and aerial buildings were selected as knowledge-unrelated items, whereas items with a large effect of building type, but with relatively few "impossible" responses, were chosen for knowledge-related items (for related items, mean effect of building type on 1–4 rating scale = 2.8, proportion of dimension responses deemed impossible = .25; for unrelated items, building type = .13, impossible = .04). Because the norming yielded only four knowledge-related dimensions, we added a fifth—type of research (deep sea or atmospheric)—which was strongly related to the category themes. The ratings that the subjects made at the end of the experiment (see the Procedure section) confirmed that this dimension was strongly thematic.

Each training example was a description of a building using all five dimensions, in random order, displayed centered on a computer screen. Table 2 shows the abstract category structure used. The assignment of abstract dimensions and, thus, of frequency to specific building features was rotated across subjects. The first items in each category, A1 and B1, were presented six times per block and so were considered HF items, in contrast to the normal low-frequency (LF) items, which appeared once per block. The atypical features in A1 and B1 (the final dimension in the table) were called *exception features*, because they were associated with the opposite category (B and A, respectively) 60% of the time, due to the HF of A1 and B1. The other features were considered *normal features* and were associated to the correct category 90% of the time.

The abstract transfer stimuli are shown in Table 2. All the transfer stimuli were presented once in each test phase. A1–A5 and B1–B5 were the trained items, A0 and B0 were novel prototype items, and AF1–AF5 and BF1–BF5 were single-feature tests.

Design. Half of the subjects were randomly assigned to the knowledge condition and saw items constructed from the features shown in the top half of Table 1, whereas the other half were assigned to the no-knowledge condition and saw items constructed from features shown in the bottom half of Table 1. The assignment of concrete stimulus dimensions to the abstract category structure was a counterbalance factor with five levels.

Procedure. The subjects were informed that they would be learning new categories but were not told about the frequency or knowledge manipulations. In order to make sure that the subjects had comparable experience with the categories, all the subjects performed five blocks of training trials, with 20 trials per block. After each block of training, the subjects were told their accuracy on that block.

On each trial, the subjects pressed a key in response to a prompt, causing an exemplar to appear on the screen. The subjects had 15 sec to decide whether the item belonged to Category Q or Category P, pressing the respective keys to indicate their choice. A "Correct" or "Incorrect" message appeared for 1.5 sec, followed by the exemplar again, with either a Q or a P on the screen to indicate the correct answer. This feedback remained visible for 4 or 8 sec to allow study,

depending on whether the subject got the trial correct or incorrect, respectively.

There were three test phases. The first phase was performed following the first block of training. The subjects were instructed to categorize the transfer stimuli (Table 2) as quickly and accurately as possible. The subjects were told to expect some new and incomplete items and to just respond as best they could. The procedure was similar to that in the training phase, with identical stimulus presentation. After the response, however, the prompt to begin the next trial was immediately displayed, without feedback. The 22 whole and single-feature items appeared in random order. Classification decision was the dependent measure for this first test phase. The second test phase was performed following the completion of the fifth and final block of training. The procedure was exactly the same as that used for the first test phase, and RT measures were also collected. For the third test phase, which immediately followed the second test, the same stimuli were used, but following categorization of each item, the subjects were asked to evaluate the typicality of the item with respect to the response category on a scale of 1 (*entirely atypical*) to 7 (*very typical*). An explanation and example of typicality was provided. The subjects were instructed to respond as accurately as possible for the third phase, without emphasizing speed, and RT measures were not collected. McDowell and Oden (1995; see also Friedman & Massaro, 1998) found no effect on categorical responses when confidence measures were also collected, so classification responses in the second and third phases should be directly comparable.

A final task was a feature-rating survey. The subjects rated each of the 20 features (i.e., both their and the alternative stimulus sets; see Table 1), indicating how predictive it was of the category themes. The instructions were, "Suppose you were trying to learn about *underwater* and *aerial* buildings in the real world (not in the context of this experiment). How useful would it be to be provided with each of the following features?" Possible responses were on a scale of 1–5, with labels of *useless*, *not very useful*, *somewhat useful*, *very useful*, and *crucial*. Since the results showed the expected effect of feature type (knowledge related or not) and a small feature frequency effect but no effect at all of the between-subjects knowledge condition, we will not discuss the survey further.

Results

The subjects in both experimental conditions learned to classify the items well. We defined a learning criterion of better than chance accuracy on LF items in the final block of training. The subjects in the knowledge condition were correct on 88% of their responses, with 2 subjects failing to reach criterion. The subjects in the no-knowledge condition were correct on 84% of their responses, with 1 subject failing to reach criterion. These 3 subjects were excluded from the analyses below.

Learning phase. Figure 1 shows training accuracy, broken down by knowledge condition and item frequency. The subjects found it considerably easier to learn frequent items, confirming a well-known prior result [repeated measures ANOVA with knowledge condition and counterbalance as between-subjects factors and blocks and item frequency as within-subjects factors, $F(1,27) = 78.92 > 4.21$, $\eta_p^2 = .75$].[1] Also, as was expected, accuracy increased with training block [$F(4,108) = 25.20 > 2.46$, $\eta_p^2 = .48$]. There was no significant main effect of the knowledge manipulation [$F(1,27) < 1$], but notably, the interaction between knowledge and frequency was significant [$F(1,27) = 6.99 > 4.21$, $\eta_p^2 = .21$]. Knowledge helped learning of LF items, but it did not seem to help learning of HF items; or alternatively, frequency had a greater effect in the no-knowledge condition.
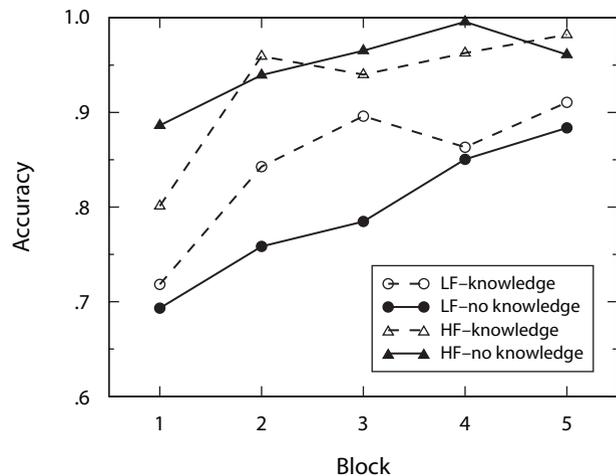

Figure 1. Experimental learning curves, Experiment 1. LF, low frequency; HF, high frequency.

Test phases. The transfer stimuli give insight into the concepts learned by the subjects in each group. As was noted above, response preferences (Tests 1–3), RT measures (Test 2), and typicality ratings (Test 3) were collected for each type of transfer item. Any RT more than two *SD*s above a subject's mean was omitted, and for the whole-item RT tests, only correct responses were included. Table 3 gives the means for each test item type, for each test, for the subjects in each of the two knowledge conditions.

Figure 2 (middle columns) shows accuracy for the training items during the three test phases. Since these items were the same as those used in training, accuracy in Test Blocks 2 and 3 would be expected to be similar to that in the last block of training (Figure 1), and it was. An ANOVA with item frequency, knowledge condition, test block, and counterbalance as factors showed that HF items were responded to more accurately than LF items [$F(1,27) = 29.31 > 4.21$, $\eta_p^2 = .52$], but this effect was reduced when concept-relevant knowledge was present [$F(1,27) = 4.81 > 4.21$, $\eta_p^2 = .15$, for the interaction]. Of course, part of this interaction may have been due to ceiling effects, since performance on HF items was above 90%. Still, the interaction was numerically obtained on all three tests and perhaps even crossed over on Test 3, supporting a knowledge-driven reduction in the frequency effect. Finally, the ANOVA showed a main effect of test number [$F(2,54) = 13.93 > 3.17$, $\eta_p^2 = .34$], reflecting the increase in accuracy with further learning. No other effects were significant, aside from a three-way interaction among test, frequency, and the counterbalance factor [$F(8,54) = 2.32 > 2.12$, $\eta_p^2 = .26$].

Figure 2 (right columns) shows the responses to the single-feature tests (see Table 2). For the purposes of analysis, correct responses were determined by *type* (ignoring frequency), not *token*, so exception features AF1 and BF1 were counted *correct* if labeled A and B, respectively. AF1 appeared in five A items and one B item, so A was deemed the correct response for the purposes of the analysis, de-
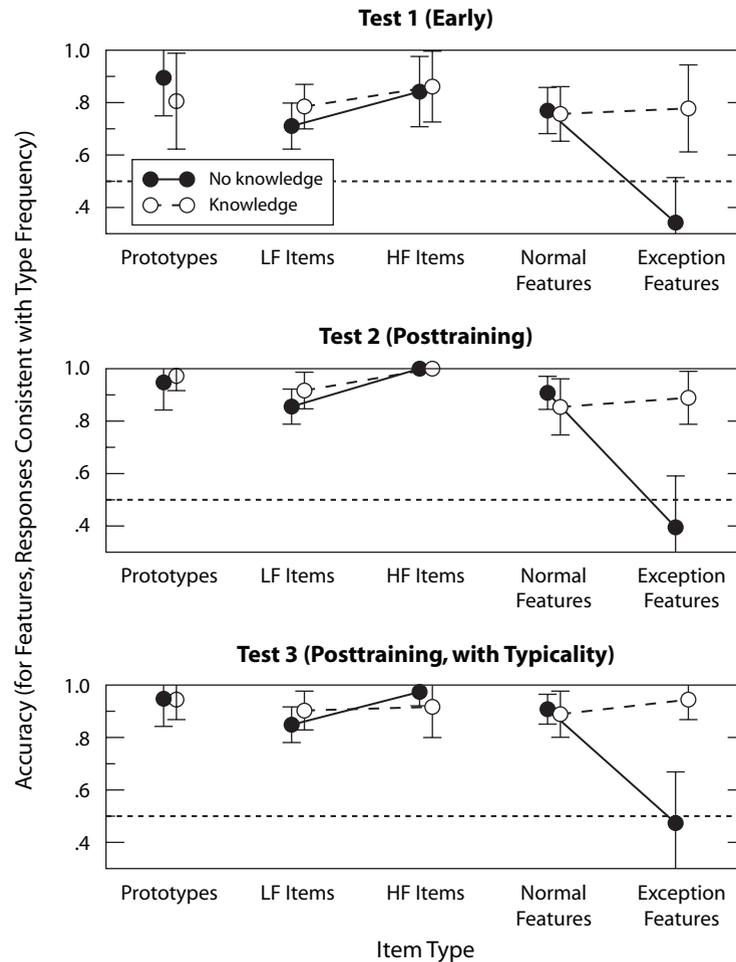
Figure 2. Experimental accuracy in Experiment 1 for prototypes, training items, and single features during tests. Test 1 was given after the first block of training, and Tests 2 and 3 were given following training. Normal items were AF2–AF5 and BF2–BF5, and exception items were AF1 and BF1. Error bars are 95% confidence intervals. LF, low frequency; HF, high frequency.

spite AF1 being associated with Category A responses only 40% of the time.[2] Given this definition of accuracy, the results show that normal features were responded to more accurately (consistently with type frequency) than were exception features [$F(1,27) = 23.48 > 4.21$, $\eta_p^2 = .47$] and that accuracy was higher for the subjects in the knowledge-related condition [$F(1,27) = 14.61 > 4.21$, $\eta_p^2 = .35$]. Critically, there was an interaction between these factors, with knowledge eliminating the tendency of the subjects to respond to exception features on the basis of token frequency [$F(1,27) = 28.58 > 4.21$, $\eta_p^2 = .51$]. Without knowledge, the subjects' choice proportions following training very nearly corresponded to the normal and exception features' token frequencies of .9 (normal) and .4 (exception). With knowledge, the subjects reversed their response preference and responded to both features consistently with their prior knowledge, apparently showing no sensitivity to frequency.

Finally, Figure 2 (left columns) shows responses to the category prototype items. Aside from a trend toward an

effect of test number [$F(2,54) = 2.69 \not> 3.17$, $\eta_p^2 = .09$], due to slightly lower accuracy on Test 1, no other effects approached significance.

In addition to knowledge's effects on response preferences, knowledge also affected how the subjects made typicality judgments. For each test item in Test 3, following collection of the response preference, we collected typicality ratings on a scale of 1–7. The raw rating was multiplied by -1 if the response preference was inconsistent with type frequency.[3] For example, if a subject classified feature AF3 as a member of Category B and gave it a typicality rating of 4, the signed typicality rating for that subject and item would be −4. Mean signed typicality ratings are shown in Table 3 and Figure 3. For the whole (training) items, HF items were viewed as more typical [$F(1,27) = 14.39 > 4.21$, $\eta_p^2 = .35$], but this frequency effect was marginally moderated with knowledge [$F(1,27) = 3.61 \not> 4.21$, $\eta_p^2 = .12$, for the interaction]. There was no main effect of knowledge condition on typicality ratings for training items [$F(1,27) < 1$].

Table 3
Test Results: Experiment 1

Condition	LF Item	HF Item	Proto.	Stimulus Feature Norm.	Stimulus Feature Except.
			Accuracy		
Test 1 (during learning)					
Knowledge	.78	.86	.81	.76	.78*
No knowledge	.71	.84	.89	.77	.34*
Test 2 (after learning)					
Knowledge	.92	1.0	.97	.85	.89*
No knowledge	.86	1.0	.95	.91	.39*
Test 3 (with ratings)					
Knowledge	.90	.92	.94	.89	.94*
No knowledge	.85	.97	.95	.91	.47*
			Reaction Times (sec)		
Test 2					
Knowledge	2.68*	2.95	2.44	1.20	1.20
No knowledge	3.64*	3.40	2.78	1.28	1.44
			Typicality Ratings (Signed)		
Test 3					
Knowledge	4.08	4.67	5.89	4.74	4.89*
No knowledge	3.63	5.50	5.42	4.95	−0.05*

Note—LF, low frequency; HF, high-frequency; Proto., prototype; Norm., normal; Except., exception. Reaction times for correct responses only are given for whole items; all responses are for single features.   *Significant simple effect of knowledge ($p < .05$, Sidak test).

For the individual features, the subjects viewed normal features as much more typical than exception features [$F(1,27) = 14.43 > 4.21$, $\eta_p^2 = .35$], but this effect essentially disappeared when the feature was related to prior knowledge [$F(1,27) = 15.47 > 4.21$, $\eta_p^2 = .36$, for the interaction]. This interaction in typicality ratings closely parallels the interaction in response preferences discussed previously, supporting the idea that knowledge can substantially overwhelm the otherwise robust effects of frequency. There was a main effect of knowledge condition as well, in which the subjects in the prior knowledge condition rated individual features as more typical more

often than did the subjects in the no-knowledge condition [$F(1,27) = 7.70 > 4.21$, $\eta_p^2 = .22$].

The RT data showed a pattern similar to that for the accuracy data. However, because of the small number of trials and fairly large variance (unsurprising, since subjects in different conditions were making judgments on the basis of lists of verbal features), few of the effects reached standard levels of significance—here or in Experiment 2. This was particularly true in the analysis of individual features, since there were very few data points for those items. In Experiment 1, the only significant effect was a simple effect of knowledge on LF items, with knowledge speeding responses [$F(1,27) = 6.08 > 4.21$, $\eta_p^2 = .18$]. Overall, although the patterns for knowledge and frequency were similar (but nonsignificant) for RTs and accuracy (Table 3), the amount of data did not allow us to draw strong conclusions on the basis of RTs.

Discussion

The results of this experiment show that exemplar frequency interacts with the presence or absence of prior knowledge. When prior knowledge was associated with the features of a category in such a way that it could be used to aid learning, the advantage of HF items over LF items was substantially reduced. Single-feature test results also dramatically shifted toward knowledge-consistent responses and away from frequency-consistent responses. The most striking result was that when frequency gave misleading evidence about a property (the exception features), learners in the no-knowledge condition classified it into the *wrong* category but those in the knowledge condition did not. Thus, our results give clear evidence that prior knowledge can reduce or even eliminate the statistical effect of exemplar frequency. This pattern of results is consistent with our hypotheses that predicted that prior knowledge would modulate empirical learning, reducing the otherwise robust empirical effects of exemplar frequency.
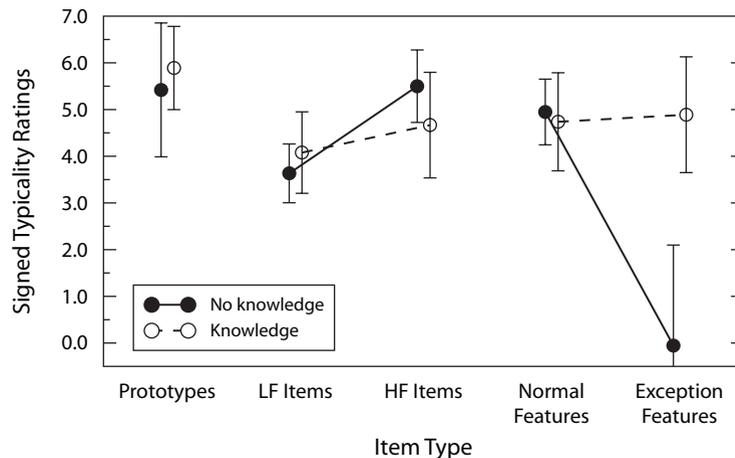

Figure 3. Typicality ratings (signed) for Experiment 1 test items (Test 3). Scores were on a scale of 1–7, multiplied by the consistency of the responses with type frequency (see the text). Error bars are 95% confidence intervals. LF, low frequency; HF, high frequency.

The results are not as clear as could be desired, however, because the interaction of frequency and knowledge possibly was influenced by a ceiling effect. The critical reversal of the subjects' categorization of the exception features cannot be explained by ceiling effects, but some of the learning accuracy and other test results could be. All of the effects for these measures take the form in which the frequency effect is reduced in the knowledge condition, where performance is very high. Although not every result seems susceptible to a ceiling effect explanation (e.g., the typicality results shown in Figure 3), we carried out another experiment that was designed to avoid ceiling effects.

EXPERIMENT 2

In Experiment 2, we made a number of small changes in procedure from Experiment 1 in an attempt to reduce ceiling effects. One change was to reduce the exemplar frequency manipulation from 6:1 to 3:1. This not only should reduce accuracy for the frequent items, but also should confirm that the effects do not depend on the presence of exception features. As will be described in detail below, the 3:1 ratio meant that exception features were now more associated to their correct category than to their incorrect category, unlike the structure used in Experiment 1. A second change was one made to the test blocks. As in Experiment 1, in Experiment 2, test blocks were given to the subjects after their first and last (fifth) blocks of training. However, only one postlearning test block was used, and each test block was identical, collecting both response preferences and typicality ratings. On the basis of the results of Experiment 1, we believed that typicality ratings would provide fine-grained information without troublesome ceiling effects.

Method

Subjects. Forty members of the New York University community received course credit for their participation. Twenty subjects were assigned to the knowledge condition, and 20 to the no-knowledge condition.

Stimuli. The stimuli in Experiment 2 were almost identical to those in Experiment 1 (Table 1). In one small change, the colonial-style and modern furniture were reversed, so that the presence of a hyphen in a feature was not a predictive cue. In another small change, the *patio* and *porch* features were changed to the more evocative *balcony* and *front porch*.

The category structure for Experiment 2 was modified so that the HF items (A1 and B1 in Table 2) were only three times more frequent than the other items. There were several consequences of this change. First, the number of trials per block was reduced from 20 to 14. The total number of training trials was thus reduced from 100 to 70, which might reduce the ceiling effect in the postlearning test. Second, the exception feature was no longer truly exceptional. Whereas in Experiment 1 the crossover features present in the HF items had been more frequently associated with the opposite category, in Experiment 2 they were merely less predictive of the correct category. Crossover features were consistent with their category in $8/14 = 57.1\%$ of the cases, whereas normal features were consistent with their category in $12/14 = 85.7\%$ of the cases.

Design. The design in Experiment 2 was identical to that in Experiment 1.

Procedure. The procedure in Experiment 2 was largely identical to that in Experiment 1, with the following small exceptions. As was noted above, blocks were now 14 trials long. RT deadlines were no longer limited to 15 sec but were unlimited. Feedback for
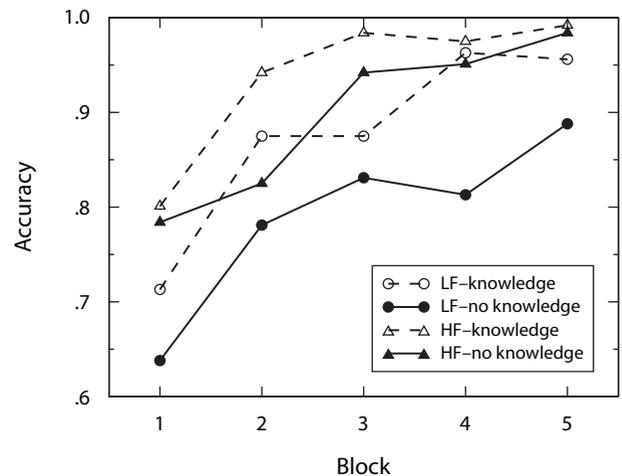

Figure 4. Experimental learning curves, Experiment 2. LF, low frequency; HF, high frequency.

correct responses was reduced from 4 to 3 sec. Both Test Phases 1 and 2 included the typicality rating task, whereas Test Phase 3 and the feature rating survey were omitted.

Results

We used the same learning criterion of better than chance accuracy on LF items in the final block of training. The subjects in the knowledge condition were correct on 96% of their responses, with all the subjects reaching criterion. The subjects in the no-knowledge condition were correct on 89% of their responses, with 1 subject failing to reach criterion. This subject was excluded from the analyses below.

Learning phase. The accuracy for each block during training is shown in Figure 4. As in Experiment 1, the subjects found it considerably easier to learn frequent items $[F(1,29) = 21.62 > 4.38, \eta_p^2 = .43]$. Also as before, accuracy increased with training block $[F(4,116) = 32.47 > 2.45, \eta_p^2 = .53]$. Unlike in Experiment 1, however, the knowledge effect was statistically reliable during learning $[F(1,29) = 5.93 > 4.38, \eta_p^2 = .17]$, whereas the interaction with frequency was not $[F(1,29) = 1.97 \ngtr 4.38]$. In this experiment, both higher frequency and prior knowledge seemed to independently increase the speed of learning.

Test phases. Response preferences, RT measures, and typicality ratings were collected for both Test 1 (after one block of training) and Test 2 (after all five blocks of training). RTs were trimmed as described above. Table 4 gives means for each test item type, for each test, for the subjects in each of the two knowledge conditions. In contrast to Experiment 1, in which the test blocks showed qualitatively similar results, the two test blocks differed substantially in Experiment 2. We will thus analyze the two blocks separately, starting with Test 2 for reasons of exposition.

Figure 5 (HF and LF items) shows accuracy for the training items during the two test phases. As before, in the test phase following training (Test 2), test accuracies were about the same as accuracies in the final block of training. HF items were responded to more accurately than were LF items $[F(1,29) = 4.74 > 4.38, \eta_p^2 = .14]$, but this effect was marginally reduced when concept-relevant knowledge

Table 4
Test Results, Experiment 2

Condition	LF Item	HF Item	Proto.	Stimulus Feature Norm.	Except.
Accuracy					
Test 1 (during learning)					
Knowledge	.76	.95*	.88	.88	.83
No knowledge	.75	.74*	.92	.77	.66
Test 2 (after learning)					
Knowledge	.97	.98	1.00	.95	.90*
No knowledge	.91	1.00	1.00	.90	.63*
Reaction Times (sec)					
Test 1 (during learning)					
Knowledge	6.73	5.72	6.21	2.84	3.35
No knowledge	6.48	6.13	5.54	3.17	3.87
Test 2 (after learning)					
Knowledge	6.20	5.65	5.29	2.24	2.34
No knowledge	6.58	5.15	6.22	2.53	2.91
Typicality Ratings (Signed)					
Test 1 (during learning)					
Knowledge	2.58	4.03	4.49	4.22	3.34
No knowledge	2.35	2.78	5.00	3.28	1.55
Test 2 (after learning)					
Knowledge	4.65	4.70*	6.20	5.32	5.03*
No knowledge	4.13	5.78*	5.70	4.75	1.11*

Note—LF, low frequency; HF, high-frequency; Proto., prototype; Norm., normal; Except., exception. Reaction times for correct responses only are given for whole items; all responses are for single features.   *Significant simple effect of knowledge ($p < .05$, Sidak test).

was present [$F(1,29) = 3.66 \not> 4.38$, $\eta_p^2 = .11$, for the interaction]. There was no main effect of knowledge in Test 2.

An analysis of the single-feature tests in Experiment 2 showed an interaction similar to that in Experiment 1 (Figure 5, normal and exception features). In Test 2, fol-

lowing learning, the subjects in the no-knowledge condition responded to normal and exception features at essentially frequency-matching rates (Ms = .90 and .63, respectively). With knowledge, however, there was little difference between normal and exception features, and responses were consistent with the category (Ms = .95 and .90). An ANOVA showed main effects of feature type [$F(1,29) = 6.74 > 4.38$, $\eta_p^2 = .19$] and knowledge [$F(1,29) = 8.16 > 4.38$, $\eta_p^2 = .22$], whereas the interaction between the two effects was marginally significant [$F(1,29) = 3.20 \not> 4.38$, $\eta_p^2 = .10$].

In Experiment 2, unlike in Experiment 1, typicality ratings were collected in both the early and the final tests. Mean signed typicality ratings, representing a more continuous measure of response preference than the simple binomial accuracy measures shown above, are shown in Table 4 and Figure 6. In Test 2, after training, for the whole-item tests, HF items were viewed as more typical than LF items [$F(1,29) = 8.37 > 4.38$, $\eta_p^2 = .22$], but this frequency effect essentially disappeared with knowledge [$F(1,29) = 7.58 > 4.38$, $\eta_p^2 = .21$, for the interaction]. Significantly, there was no ceiling effect in the comparison, and the effect of knowledge was to *reduce* the typicality of HF items. Thus, the effect cannot be explained by a high level of responding across the board by the subjects in the knowledge condition. There was no main effect of knowledge condition on typicality ratings for training items [$F(1,29) < 1$].

After five blocks of learning in Experiment 2, the subjects rated individual features much as they did in Experiment 1. As is shown in Figure 6, normal features were rated as more typical than were exception features
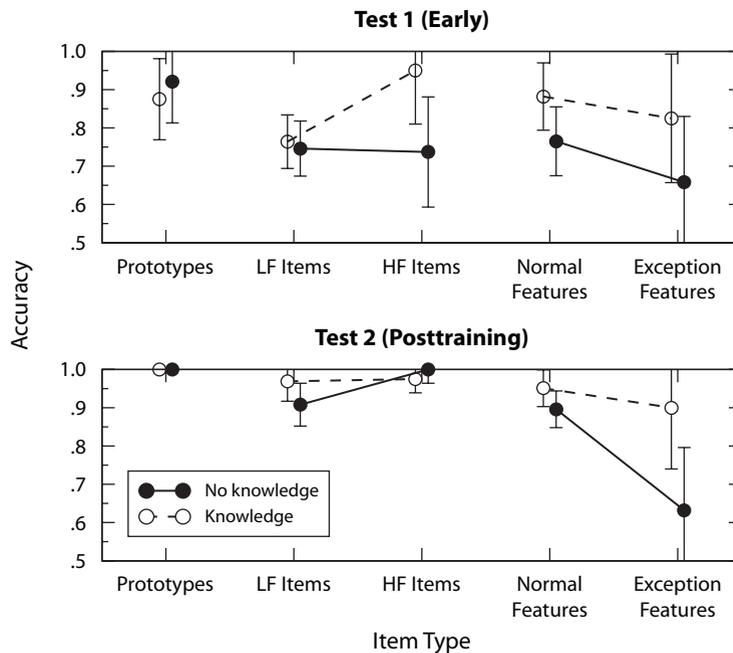

Figure 5. Experimental accuracy in Experiment 2 for training items and single features during tests. Test 1 was given after the first block of training, and Test 2 was given following training. Error bars are 95% confidence intervals. LF, low frequency; HF, high frequency.
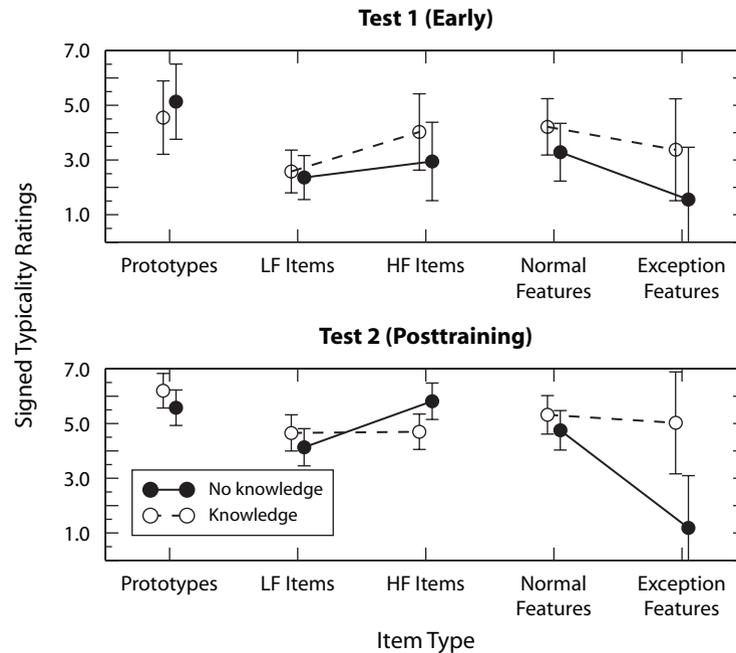
Figure 6. Typicality ratings (signed) for Experiment 2 test items. Scores were on a scale of 1–7, multiplied by the consistency of the responses with type frequency (see the text). Error bars are 95% confidence intervals. LF, low frequency; HF, high frequency.

$[F(1,29) = 7.30 > 4.38, \eta_p^2 = .20]$, but this effect essentially disappeared when the feature was related to prior knowledge $[F(1,29) = 5.29 > 4.38, \eta_p^2 = .15]$. When prior knowledge was applicable, all the features seemed more typical $[F(1,29) = 8.94 > 4.38, \eta_p^2 = .24]$.

The results above from the postlearning test phase in Experiment 2 closely resemble the results from Experiment 1. However, the pattern following just one block of training was strikingly different. Rather than knowledge's reducing the strength of the frequency effect, in Test 1 of Experiment 2 knowledge dramatically *increased* the strength of the frequency effect (Figure 5, top center). There was no reliable main effect of frequency $[F(1,29) = 2.88 \not> 4.38]$, but there was a main effect of knowledge condition $[F(1,29) = 4.68 > 4.38, \eta_p^2 = .14]$ and a marginal interaction between knowledge and frequency $[F(1,29) = 3.47 \not> 4.38, \eta_p^2 = .11]$. Comparing the HF items only, accuracy without knowledge ($M = .74$) was significantly lower than accuracy with knowledge ($M = .95$) $[F(1,29) = 5.18 > 4.38, \eta_p^2 = .15]$. As for accuracy on single-feature tests, whereas the postlearning test showed an interaction between feature type and knowledge, no such result was seen in Test 1 (Figure 5, upper right). There was a marginal effect of knowledge $[F(1,29) = 3.84 \not> 4.38, \eta_p^2 = .12]$, only a weak trend toward an effect of feature type $[F(1,29) = 2.04 \not> 4.38, \eta_p^2 = .07]$, and barely a hint of an interaction $[F(1,29) = 0.25]$. As will be discussed below, these results seem to suggest that frequency may not play the same role after just one block of training that it does after five blocks.

The typicality ratings likewise showed different patterns early and late in training. Whereas late tests of the

trained items showed a reduced frequency effect with knowledge (Figure 6, bottom, LF and HF items), the early tests showed no such pattern (Figure 6, top). There was a marginal effect of item frequency $[F(1,29) = 4.09 \not> 4.38, \eta_p^2 = .12]$, a very weak trend toward an effect of knowledge condition $[F(1,29) = 1.68 \not> 4.38]$, and no interaction $[F(1,29) < 1]$. We observed a similar difference in the typicality ratings for the single features, with a large interaction following training (Figure 6, bottom, normal and exception features), but no interaction early in training (Figure 6, top). There were marginal effects of feature type $[F(1,29) = 3.86 \not> 4.38, \eta_p^2 = .12]$ and knowledge $[F(1,29) = 2.81 \not> 4.38, \eta_p^2 = .09]$, but no interaction between the two factors $[F(1,29) < 1]$. Once again, the early test results are inconsistent with the later test results.

Discussion

The postlearning results of Experiment 2 support our conclusions from Experiment 1. Prior knowledge reduces, and in some cases eliminates, the effects of frequency on both well-trained items and single features of those items. For whole-item response accuracy, a substantial frequency effect was eliminated when knowledge could be used. Even more striking, when typicality ratings were used to avoid ceiling effects in accuracy, the same effect occurred, with a large difference in typicality ratings without knowledge becoming no difference at all with knowledge. In another substantial effect, a parallel to the reversal of response preference to exception features found in Experiment 1 was observed here. Without knowledge, the subjects seemed to frequency-match single-feature tests, but with knowledge, frequency had very little effect.

This consistency was not observed during the early stages of learning. After a single block of training, most of the observed effects were not yet evident and, in fact, seemed often to be reversed. For example, whereas well-trained item typicality ratings were sensitive to frequency without knowledge (change in ratings, $d = 1.8$) and insensitive to frequency with knowledge ($d = 0.04$), weakly trained item typicality was *insensitive* to frequency without knowledge ($d = 0.1$) but *quite sensitive* to frequency with knowledge ($d = 1.4$). In Experiment 1, the patterns of results for the different test stages were broadly consistent. Why was the first test block different here?

One factor is that the test in the present experiment was, in fact, earlier than that in Experiment 1. Recall that we reduced the frequency manipulation in Experiment 2 in order to reduce ceiling effects. Thus, by reducing frequency, we also reduced the number of trials in each block, and therefore, Test 1 occurred after only 14 items had been viewed.

As the modeling results will reveal (see the next section), one explanation of the ultimate pattern of results is that prior knowledge serves to reduce frequency effects by boosting the activation of thematically consistent features and reducing the activation of inconsistent features, so that infrequent features receive some activation even when they are not presented. Of course, such an effect can take place only once the thematic knowledge that relates the features is detected. Thus, if some of the subjects have not identified what the category themes are after 14 trials, they cannot reveal the predicted interaction. In addition, subjects must identify the varying dimensions of the stimuli and figure out the pairs of features for each dimension, a process that may be facilitated by prior knowledge. Any hypotheses about these early stages of category learning are necessarily quite speculative, however, since little is known about the processes by which representations of this sort of stimuli are initially formed. Further research will be necessary to understand how knowledge and frequency interact at the early stages of category learning.

MODELING AND ANALYSIS: KRES MODEL

The KRES model of category learning was designed to account for a wide variety of category learning and categorization data. In particular, it is one of only a few computational models of category learning that can take into account the effects of prior knowledge (see also Heit & Bott, 2000). KRES is an interactive activation model of categorization (McClelland & Rumelhart, 1981), with error-driven training by contrastive Hebbian learning (O'Reilly, 1986) and prototype-like representations of categories (although see Harris & Rehder, 2006, for an exemplar-based variation of KRES). The model has been used to account for how knowledge affects learning rate, RTs, the classification of features not related to knowledge, and the integration of conflicting prior and empirical knowledge (Rehder & Murphy, 2003).

The interactive activation properties of KRES give it very different dynamics from typical connectionist networks used to model category learning. Each node is connected to other nodes by bidirectional connections, which slowly cause changes in activation over many time steps. The stimulus representation is *added* to the activation of the input nodes, so other influences can change or even override the presented values. For example, if no information about a dimension is provided, the features of that dimension (usually two) initially have equal, moderate activation. However, top-down influences from activated category response nodes can cause those nodes to become increasingly or decreasingly activated, so as to be most consistent with the experience represented by the weights. Other influences on activations are fundamental properties of KRES networks.

Importantly, KRES can implement prior knowledge in different ways. In one way (see Figure 7, left), which we call KRES/F to indicate a path between features and category nodes, prior knowledge is represented by lateral excitatory connections among features of the input represen-
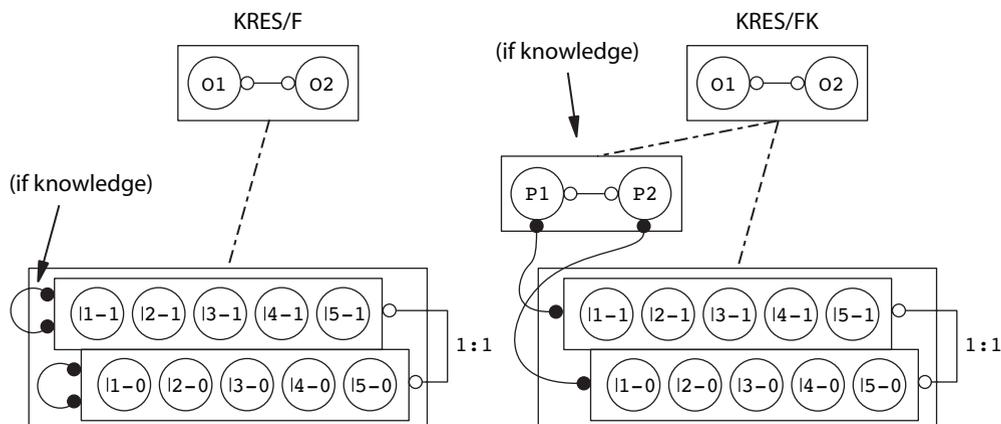

Figure 7. KRES/F model used in simulations and KRES/FK model. There are fixed inhibitory connections (open circles) between each pair of input nodes (I), between the two output nodes (O), and between the two prior knowledge nodes (P; KRES/FK only). With knowledge, there are fixed excitatory connections (filled circles) among all related elements of an input layer (KRES/F) or between input nodes and a prior knowledge node (KRES/FK). All the input nodes are connected to both output nodes with trainable connections (dotted line), as are the two prior knowledge nodes (KRES/FK). Without knowledge, the models are identical.

tation. Features that are consistent with prior knowledge spread their activation to other related features, increasing their activation. (These connections are assumed to represent previously learned conditional frequencies, or prior instructions, or other sorts of information in long-term knowledge.) The KRES/F model is a single-route model, in which knowledge affects representations and learning but does not have independent associations with responses.

Alternatively, KRES can have prior concept nodes that are connected to the input features and that, with training, become associated directly with response nodes (Heit & Bott, 2000). In this model, which we call KRES/FK to indicate both feature and knowledge node connections to concept nodes, there are two routes to response selection (Figure 7, right). Since our earlier work has suggested that the type of prior concept may determine the specific way that knowledge affects learning (Harris & Rehder, 2006), we have chosen for this project to focus on the KRES/F model. The knowledge of aerial and underwater buildings used in the experiment is not in the form of prior concepts but, instead, seems to involve interrelations among stimulus features, reflecting knowledge that, for example, astronauts are more likely to perform atmospheric research than deep-sea research. In fact, there *are* no existing categories of underwater and aerial buildings with the features we have attributed to them.

Finally, we note that as a prototype model, with no representation of exemplars, the KRES/F model treats each input as an independent novel stimulus and represents frequency only as learning-induced differential patterns in its weights.

Experiment 1

To examine the behavior of the model, we first set it up to learn the category structure of Experiment 1, under a learning process that was the same as the one that experimental subjects followed. The model has four free parameters: learning rate (l.r.), a measure of node response sharpness called α (fixed at 1.0 in Rehder & Murphy, 2003, but varied here), and the strength of the fixed inhibitory and excitatory connections (inhib. wt. and excit. wt., respectively). Larger values of α force activations to be nearer 0 or 1, larger values of inhibitory connections push pairs of nodes to have more nearly opposite activations, and larger values of excitatory connections push knowledge-related nodes to be more similar in activation. We performed a parameter space survey over a region of the parameter space that produced nondegenerate results. For each of the 720 parameter settings sampled, the average accuracy of the model (over five replications) was computed and compared with the empirical results (combining Test 2 and Test 3 data) using a root-mean-squared scaled deviation (RMSSD) goodness-of-fit metric (Schunn & Wallach, 2005). The RMSSD metric indicates the mean squared deviation between the model and the data, in units of standard errors of each data point. RMSSD prefers tighter fits to measures with small standard error and allows weaker fits to measures with large standard error.

Figure 8 (left) shows the results of the best-fitting model, overlaid on top of the empirical accuracy data. The model was successful at fitting the 10 data points with four parameters, usually within the 67% confidence intervals of the empirical data (plotted), and in all but one case within the 95% confidence intervals. The RMSSD was 0.69, with l.r. = 0.05, α = 1.5, inhib. wt. = −1.5, and excit. wt. = 0.125. (Since a grid search was used, rather than a parameter optimization approach, slightly better fits are likely possible.) In addition to the strong quantitative fit, KRES shows the following empirically observed qualitative properties: (1) higher accuracy on HF items than on LF items, which (2) is reduced with knowledge; (3) high accuracy on novel prototype items; and (4) a very large difference between normal and exception single features without knowledge, which (5) nearly disappears in the presence of prior knowledge.

The key empirical result was the interaction between knowledge and frequency (Figure 2). In the experiment, frequency increased performance without knowledge, but not when knowledge was present. Likewise, the normal features showed a relatively small classification change with knowledge, whereas the exception features showed a very large change. The best fit of the model likewise showed an increase with knowledge in LF accuracy (from .87 to .90) and a smaller change in HF accuracy (from .96 to .97), as well as a relatively small change in normal feature classification (from .86 to .92), as compared with the huge knowledge effect on exception feature responses (from .43 to .89).

Although a good fit such as this one is compelling support for a model, a model that can fit any arbitrary pattern of data by changes in parameter settings says relatively little about the underlying processes. Recent work has argued that a successful model should have, in addition to a good quantitative fit, a qualitative fit that is based on the model's architecture, rather than on carefully tuned parameters (Pitt, Kim, Navarro, & Myung, 2006; Pitt & Myung, 2002).

Here, one of the most important qualitative patterns was the reduction in the magnitude of the frequency effect when knowledge was present. Figure 9 (left) shows a scatter graph of the frequency effect on whole items (HF accuracy minus LF accuracy), with and without knowledge, across the parameter space of the model, along with the empirical result. Across its parameter space, KRES shows an interaction in which the frequency effect is larger without knowledge than with knowledge. This pattern matches the result found in our experiments. Figure 9 (right) shows an analogous graph for the single-feature tests. The model tends to show very large differences for different feature types (normal feature accuracy minus exception feature accuracy) without knowledge but much less or no difference with knowledge, matching the empirical result.[4]

Figure 9 shows that the model tends (in nondegenerate areas of the parameter space) to account for the qualitative empirical effects and does not account for a number of possible empirical effects that could have been observed but were not. The model is not so flexible that it can account for any arbitrary pattern of empirical results simply by a change in parameters. Its good quantitative fit thus supports the model's architecture as an explanation of the
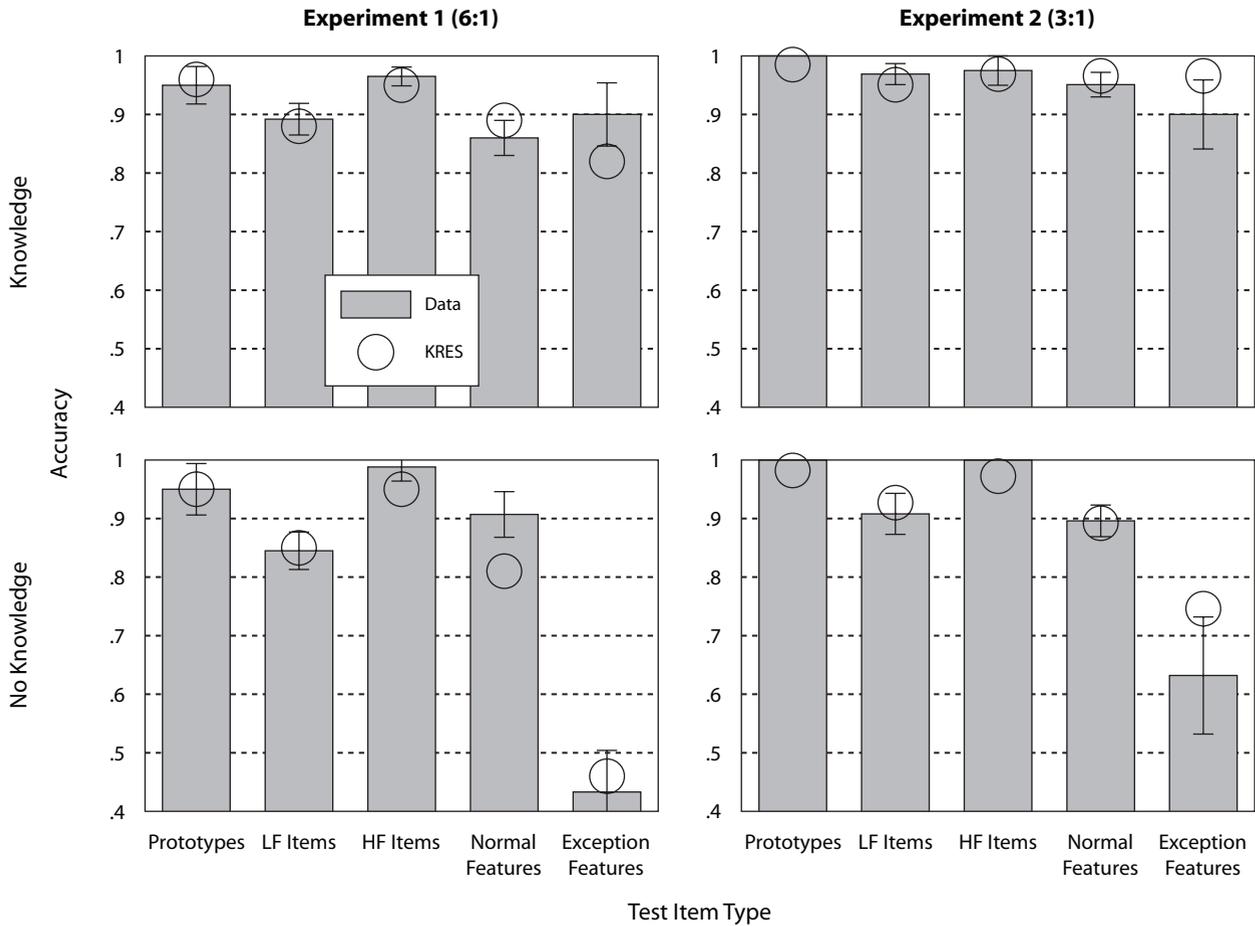
Figure 8. Best fit of the KRES model to the data from Experiments 1 (Tests 2 and 3) and 2 (Test 2). Error bars (standard errors of the means, or 67% confidence intervals) are shown for the empirical data. LF, low frequency; HF, high frequency.
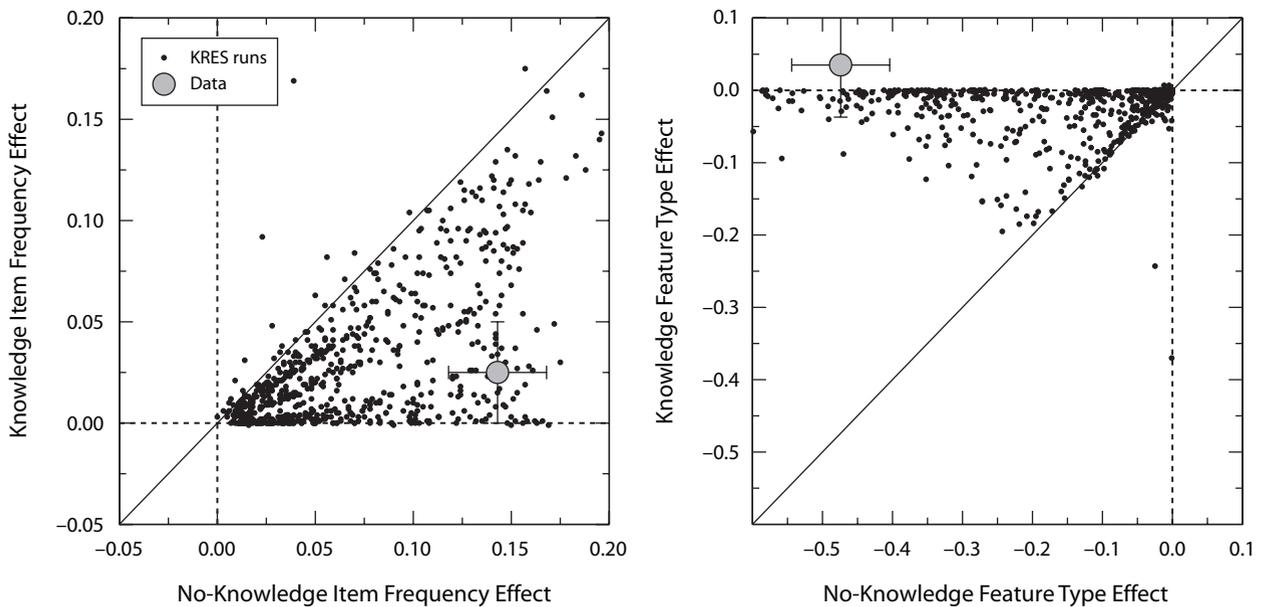

Figure 9. Scatter graph of the magnitude of the effect of item frequency and single-feature type on response preferences in the knowledge and no-knowledge conditions in Experiment 1 for the KRES model over a wide range of parameter settings. Error bars (standard error of the mean, or 67% confidence intervals) are shown for the empirical data.

processes being investigated by the experiment, as will be discussed further below.

Experiment 2

To further investigate the performance of the KRES model on this task, we fit the model to the task in Experiment 2. For the purposes of modeling, the only difference between the two experiments was that the strength of the frequency variation changed from 6:1 to 3:1. Otherwise, the procedure was the same. The model was run over a wide variety of parameter settings, to get a qualitative understanding of the model's performance, and the RMSSD metric was used to find parameters that fit the Test 2 results well.

The best result from this coarse fitting procedure was found with the parameters l.r. $= .03$, $\alpha = 3$, inhib. wt. $= -0.7$, and excit. wt. $= 0.3$. The RMSSD measure was 0.84, indicating that the average error was less than the *SEM* of the data. Following all five blocks of training, each of the major qualitative patterns seen in Table 4 was observed in the KRES model (see Figure 8, right). The model showed a frequency effect on trained items that was reduced in the presence of prior knowledge (from $.97 - .93 = .04$ to $.97 - .95 = .02$) and, likewise, showed the reduced effect of feature type on the single-feature tests, when prior knowledge was present (from $.89 - .75 = .14$ to $.97 - .97 = 0$). These results confirm that KRES can account for the interacting effects of frequency and prior knowledge over a range of frequency differences.

The results of simulating the first test, following the first block of training, were somewhat different. Recall from above (and Figure 5) that early in training, we found an effect of frequency on trained items only with knowledge, not without knowledge, contrary to the Test 2 effects. The KRES model does not show this result, instead just showing the same pattern as the late test, but with lower accuracy overall. For example, whereas our subjects responded equally accurately to LF and HF trained items without knowledge, KRES showed a frequency advantage. In addition, knowledge increased both HF and LF response accuracies in KRES by roughly equal amounts. The model's performance on the single-feature tests also differed from our data. Although we observed single-feature tests to be nearly as accurate as tests of the trained items, the model was considerably less accurate on single-feature tests.

This pattern of results, with the model qualitatively fitting relatively well late in training but very poorly early in training, suggests that KRES does not capture all aspects of the learning process, and especially not the initial phases of learning. It may be, for example, that people's representations of dimensions, features, and stimuli are not coherent early in learning, which KRES, with its hand-constructed representations, cannot capture. However, KRES's success in quantitatively fitting Experiment 1's test results and quantitatively fitting the late phase of Experiment 2 does suggest that an understanding of the model can provide some insight into the processes that may lead to the knowledge–frequency interactions.

Analysis

The KRES model was able to account for the data by using bidirectional connections, error-driven learning, and knowledge represented as lateral connections among features. The basic frequency effect—faster learning and more accurate responses to HF items—appears to be a result of error-driven learning: The weights between input features and category labels are more frequently updated by errors on HF items than by errors on LF items, pulling the prototype represented by the weights toward the HF items (Barsalou et al., 1998). The high test accuracy on the (untrained) prototype items is due to the prototype architecture of the model. The accuracy on single-feature tests likewise follows from the prototype representations of the model, with the normal features having strong associations with the category labels but the exception feature having only weak associations.

When knowledge is added, the picture becomes more complicated. The knowledge–frequency interaction could be due to several aspects of the model, including changes in activation due to lateral connections, changes in activation due to top-down feedback, potential dynamic processes in activation, and changes in weights due to shifts in learning. It is important to understand what aspect of the model yields the observed behavior. To address this, we consider the activations of the input nodes, which (through recurrent connections and KRES's constraint satisfaction processes) represent the combined influence of all aspects of the model.

Figure 10 illustrates the input nodes of a typical model with the best-fit parameters for Experiment 1, following training. Each rectangle represents a node's activation when presented with the specified input pattern. For example, the upper left boxes represent the five "1" input nodes (I1-1 to I5-1 in Figure 7; the five "0" input nodes are not shown), when presented with the 11110 pattern (the HF item). The rightmost box is the exception feature. The height of each box indicates the activation of that input node once the network has settled into a steady state. The width of each box represents the learned weight between that node and the correct category response node. The area of each box thus represents the signal that the input node provides to the response node. The total area, which is the total weighted input to the category node from these units, is in the column on the right.

The key result of this visualization is that the total amount of activation provided to the category node is determined mostly by the weights and is only minimally determined by changes in activation. Consider the first two rows of the no-knowledge column. Since the exception feature is only weakly associated with the response, the value of that feature has only a weak effect on categorization. However, when knowledge is available (left column), the weight is considerably larger, and the change in total weighted activation is much less. The change in input node activation (height) due to knowledge is almost imperceptible and has little effect on the response. A similar pattern occurs with single-feature tests (bottom two rows). Without knowledge, the exception feature is weakly weighted and contributes little to the category
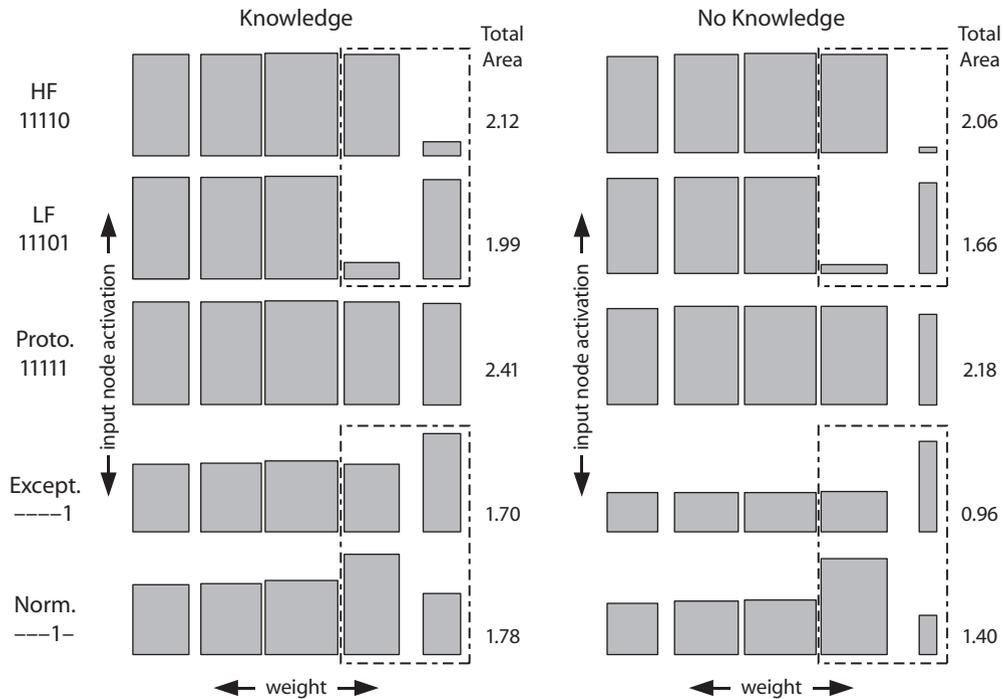
Figure 10. Activation of KRES input nodes, following learning, for a typical run of the model for Experiment 1. The first three rows represent whole-test items; the bottom two rows represent single-feature tests. Each rectangle represents a "1" input node, with the height proportional to the equilibrium activation of the node and the width proportional to the weight between the node and the correct response node. The area of each rectangle thus represents the contribution of the input node to the response node's activation. The number to the right is the sum of the areas, is equal to the input nodes' contribution to response node activation, and is related to response probability. Key comparisons are outlined. HF, high frequency; LF, low frequency; Proto., prototype; Except., exception; Norm., normal.

node activation. With knowledge, the exception feature is weighted more strongly, and there is little difference in total weighted activation. Note that here the effects of lateral and top-down feedback are more prominent. Without knowledge, top-down feedback yields higher activation of the missing features in the normal feature case, which in turn yields higher overall weighted activation. With knowledge, this same effect occurs to some extent, but now both exception and normal features have higher activation, due to the lateral connections.

From this visualization and analysis, we can conclude that the effects of knowledge on test accuracy are due primarily to different patterns of learned weights and only secondarily to test time effects of resonance and constraint satisfaction. But why are the exception features weighted more strongly when knowledge is present in the network? The answer here does have to do with issues of resonance and constraint satisfaction. In the early stages of learning, the lateral excitatory knowledge connections tend to increase the activation of all the input nodes when other input nodes are active. The exception feature in a stimulus like 11110 tends to be more strongly activated with knowledge (i.e., rather than 0, the last dimension takes on a positive value), which results in larger weight changes. The increase in activation can be seen in the top row of Figure 10, where the height of the rightmost rectangle is considerably greater with knowledge. The effect on over-

all weighted activation is minimal, but the effect on learning, across many trials, is significantly more substantial.

GENERAL DISCUSSION

The experimental results described here show a strong interaction between an important structural property of a category, exemplar frequency, and an important external factor in concept learning, the content of the categories. When thematic prior knowledge was relevant to the concept being learned, the learning advantage for HF items over LF items was greatly reduced. Likewise, in postlearning tests, knowledge improved classification of the LF items, whereas it had little effect on HF items. Knowledge never completely eliminated the frequency effects on the trained items, however, suggesting that categorization decisions are influenced by both sorts of information. An analogous pattern of results was seen with the single-feature tests. Without knowledge, the subjects frequency-matched associations between features and category responses, but with prior knowledge, responses become consistent with that knowledge and inconsistent with frequency.

One important sign of progress in the field of category learning has been the creation of a number of explicit computational models that account for a wide variety of empirical data. However, until recently, there have been

few models that account for the effects of prior knowledge on category learning. The KRES class of models is one exception, and a notable result from the present study is KRES/F's ability to account for the learning data presented here. This success arose because the KRES architecture allows both prior knowledge and empirical information to make their influence felt. Whereas some other models of prior knowledge (e.g., Pazzani's [1991] PostHoc model) have viewed prior knowledge as biases over a set of candidate rules, the data clearly require a model that can represent probabilistic associations of features with category labels. KRES/F naturally represents this sort of frequency information while, at the same time, accounting for the profound effect that prior knowledge can have on category learning.

Further research is required to identify more precisely the representations of the prior knowledge involved in category learning and the nature of the interaction between that knowledge and the regularities inherent in observed category members. In this regard, we find Heit's (1994) distinction between *distortion* and *integration* models helpful in delineating the space of possibilities. In a distortion model, prior knowledge works to alter, or distort, a learner's representation of an input stimulus. In an integration model, prior knowledge and an empirical learning component make independent contributions to the categorization decision. In this light, KRES/F can be considered a kind of distortion model, because prior knowledge works (via constraint satisfaction) to rerepresent the input in a manner that is more consistent with prior knowledge. Lateral knowledge connections in the model alter the stimulus representations during processing, strengthening the activation of knowledge-consistent feature nodes and weakening the activation of knowledge-inconsistent features. These changed representations yield changed association weights, which strongly affect response patterns. In contrast, the other variant of KRES we described, KRES/FK (see Figure 7), acts more like an integration model, because the prior knowledge nodes and the feature nodes have their own, mostly independent, influence on the category labels. (Another example of an integration model is Heit & Bott's [2000] Baywatch model, a feedforward network that incorporates connections to category label nodes from both prior knowledge nodes and feature nodes.)

It may be, however, that neither a distortion model nor an integration model is correct in an absolute sense but, rather, the appropriate model depends on the type of prior knowledge involved. For example, our decision to model the present data with KRES/F was based on prior experimental work demonstrating that the category themes (underwater and aerial buildings) did not correspond to concepts that were already familiar to university students (Murphy & Allopenna, 1994). Although our modeling effort was successful, the aim of future tests of models could be to show that only one particular way of combining prior knowledge with empirical information (integration, distortion, or yet some other possibility) can adequately account for the observed learning performance. The category structure tested in the present study was not designed to discriminate between different classes of models (e.g., distortion vs. integration). A complete test would require comparison of different kinds of knowledge—thematic versus preexisting concepts.

Progress is already being made in conducting more sensitive tests. For example, testing categories that most people were familiar with (shy person, frequent traveler, college graduate, etc.), Heit (1994) found that an integration model provided the best account (although see also Heit, 1998). Similarly, Harris and Rehder (2006) found that an integration model (specifically, a version of KRES elaborated with exemplar nodes) provided a better account of learning both linearly and nonlinearly separable categories that corresponded to familiar concepts (Wattenmaker et al., 1986). We expect that these and new studies, combined with model-testing methodology that has been applied successfully in the past, will shed new light on the details of how prior knowledge influences category learning.

By showing how the structural effects of new categories and prior knowledge related to the content of those categories interact, we have supported a view of category learning in which prior knowledge can be a complex and nontrivial factor. By using the KRES model of category learning to account for the experimental data, we have made progress in understanding how those complexities might be realized and how the old and new representations involved in categorization and category learning affect each other.

REFERENCES

Barsalou, L. W. (1985). Ideals, central tendency, and frequency of instantiation as determinants of graded structure in categories. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **11**, 629-654.

Barsalou, L. W., Huttenlocher, J., & Lamberts, K. (1998). Basing categorization on individuals and events. *Cognitive Psychology*, **36**, 203-272.

Friedman, D., & Massaro, D. W. (1998). Understanding variability in binary and continuous choice. *Psychonomic Bulletin & Review*, **5**, 370-389.

Harris, H. D., & Rehder, B. (2006). Modeling category learning with exemplars and prior knowledge. In R. Sun & N. Miyake (Eds.), *Proceedings of the 28th Annual Conference of the Cognitive Science Society* (pp. 1440-1445). Mahwah, NJ: Erlbaum.

Heit, E. (1994). Models of the effects of prior knowledge on category learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **20**, 1264-1282.

Heit, E. (1998). Influences of prior knowledge on selective weighting of category members. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **24**, 712-731.

Heit, E., & Bott, L. (2000). Knowledge selection in category learning. In D. L. Medin (Ed.), *Psychology of learning and motivation* (Vol. 39, pp. 163-199). San Diego: Academic Press.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: Part I. An account of basic findings. *Psychological Review*, **86**, 375-407.

McDowell, B. D., & Oden, G. C. (1995). *Categorical decision, rating judgments, and information preservation*. Unpublished manuscript, University of Iowa.

Mervis, C. B., Catlin, J., & Rosch, E. (1976). Relationships among goodness-of-exemplar, category norms, and word frequency. *Bulletin of the Psychonomic Society*, **7**, 283-284.

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

Murphy, G. L., & Kaplan, A. S. (2000). Feature distribution and background knowledge in category learning, *Quarterly Journal of Experimental Psychology*, **53A**, 962-982.

Nosofsky, R. M. (1988), Similarity, frequency, and category representations. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **14**, 54-65.

Novick, L. R. (2003). At the forefront of thought: The effect of media exposure on airplane typicality. *Psychonomic Bulletin & Review*, **10**, 971-974.

O'Reilly, R. C. (1986). Biologically plausible error-driven learning using local activation differences: The generalized recirculation algorithm. *Neural Computation*, **8**, 895-938.

Pazzani, M. J. (1991). Influence of prior knowledge on concept acquisition: Experimental and computational results. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **17**, 416-432.

Pitt, M. A., Kim, W., Navarro, D. J., & Myung, J. I. (2006). Global model analysis by parameter space partitioning, *Psychological Review*, **113**, 57-83.

Pitt, M. A., & Myung, I. J. (2002). When a good fit can be bad. *Trends in Cognitive Sciences*, **6**, 421-425.

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

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

Schunn, C. D., & Wallach, D. (2005). Evaluating goodness-of-fit in comparison of models to data. In W. Tack (Ed.), *Psychologie der Kognition: Reden and Vorträge anlässlich der Emeritierung von Werner Tack* (pp. 115-154). Saarbrücken, Germany: University of Saarland Press.

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.

Wattenmaker, W. D., Dewey, G. I., Murphy, T. D., & Medin, D. L. (1986). Linear separability and concept learning: Context, relation properties, and concept naturalness. *Cognitive Psychology*, **18**, 158-194.

NOTES

1. We will report the results of statistical tests by comparing the statistic in question with the critical value of that statistic, assuming $p = .05$, and by providing the partial eta-squared (η_p^2) measure of effect size.

2. Coding accuracy by token frequency would result in the same interaction, only the knowledge group would then have lower accuracy on the exception items.

3. Signed typicality ratings are more appropriate than typicality ratings that ignore the classification. A rating of 3 is very different depending on whether a subject has classified an item into the correct versus the incorrect category. Someone who rates an item as a 1 in the incorrect category is "more correct" than someone who rates it as a 7, which is reflected in -1 being a higher score than -7.

4. A parallel analysis is to compare the effects of knowledge for each type of test. Like the data, KRES tends to show a positive knowledge effect on LF items, a weaker or nonexistent knowledge effect on HF items, a very strong knowledge effect on exception features, and a weaker or nonexistent knowledge effect on normal features.