

Prior knowledge enhances the category dimensionality effect

AARON B. HOFFMAN, HARLAN D. HARRIS, AND GREGORY L. MURPHY
New York University, New York, New York

A study of the combined influence of prior knowledge and stimulus dimensionality on category learning was conducted. Subjects learned category structures with the same number of necessary dimensions but with more or fewer additional, redundant dimensions and with either knowledge-related or knowledge-unrelated features. Minimal-learning models predict that all subjects, regardless of condition, either should learn the same number of dimensions or should respond more slowly to each dimension. Despite similar learning rates and response times, subjects learned more features in the high-dimensional than in the low-dimensional condition. Furthermore, prior knowledge interacted with dimensionality, increasing what was learned, especially in the high-dimensional case. A second experiment confirmed that the participants did, in fact, learn more features during the training phase, rather than simply inferring them at test. These effects can be explained by direct associations among features (representing prior knowledge), combined with feedback between features and the category label, as was shown by simulations of the *knowledge resonance*, or KRES, model of category learning.

Most category learning involves an interplay between prior knowledge and basic associative learning. When we learn about a new kind of thing, we generally know something about that domain (animals, electronic devices, medical conditions) already. But learning the new concept nevertheless requires that we encounter examples and associate properties to a category representation. In the present article, we investigate how the creation of these associations interacts with prior knowledge during learning.

Relevant prior knowledge speeds concept learning (Murphy & Allopenna, 1994; Pazzani, 1991). For instance, an experienced game-player's knowledge of past games and their rules makes it easy to learn a novel game, and a car mechanic's knowledge of engines helps him or her quickly learn about a new car model. Even a minimal amount of prior knowledge (only one in five features connected to prior knowledge) can reduce the number of learning trials and increase what is learned about novel concepts (Kaplan & Murphy, 2000).

However powerful, the effects of prior knowledge are not addressed by most current categorization theories. Although our understanding of concept learning has advanced over the past 30 years through the development of sophisticated models, such models address how people learn to respond to category exemplars varying on simple dimensions (e.g., line length, orientation, color), with little or no connection to prior conceptual knowledge. The models do well in characterizing people's behavior in learning these types of simple concepts. In the next sections, we will discuss the formal properties of some of these models and then relate those properties to the issue of how knowledge influences category learning.

Minimal-Learning Models

Standard category-learning models are based on error-driven learning mechanisms, whether they are exemplar (Kruschke, 1992; Kruschke & Johansen, 1999), rule-based (Nosofsky, Palmeri, & McKinley, 1994), or hybrid (Ashby, Alfonso-Reese, Turken, & Waldron, 1998; Erickson & Kruschke, 1998, 2002) models. Such models assume that category learning is driven by the goal of reducing classification error to zero. This assumption derives from earlier work on associative learning, such as the Rescorla–Wagner learning model (Rescorla & Wagner, 1972), which proposes that once a stimulus can be predicted by other stimuli, no further associative learning will take place. In the classical conditioning paradigm of *blocking* (Kamin, 1969), once an animal has learned that a light predicts a shock and is then exposed to stimuli consisting of the light plus a tone, resulting in shock, it will not learn that the tone predicts the shock. The shock is already fully predicted by the light, and therefore, there is no prediction error to drive the formation of an association between the tone and the shock.

In many models of human category learning, forming associations (between features or exemplars and a category) works in a similar way. When the model makes incorrect classifications, associations are changed to decrease this error. When the category is correctly predicted, this updating stops, and no further learning occurs. In computational models, this assumption can be found in learning algorithms using versions of the Widrow–Hoff learning rule (Widrow & Hoff, 1960), such as backpropagation (for a discussion, see Gluck & Bower, 1988; Kruschke, 1993). When models use error in this way, loosely speaking, only

A. B. Hoffman, aaron.hoffman@mail.utexas.edu

the minimum amount necessary to achieve correct classification will be learned. Recall that in the case of blocking in classical conditioning, once the animal has learned that the light predicts the shock, it does not learn that a tone also predicts the shock. In category learning, the analogy would be that once people have learned that some stimulus properties determine categorization, other properties will not be learned. We call such models *minimal-learning models*, in that they learn only what is necessary to classify correctly. Not all error-driven models are minimal-learning models, because they include additional mechanisms for learning more than what the error signal might allow. For example, SUSTAIN (Love, Medin, & Gureckis, 2004) has an unsupervised learning component that changes its cluster representation of the category when it encounters unusual exemplars. As we will demonstrate later, KRES (Rehder & Murphy, 2003) uses interactive activation to learn more than the minimum necessary.

In previous work (Hoffman & Murphy, 2006, Experiment 3), we investigated whether people are minimal learners by examining category learning of objects with fewer (four) or more (eight) stimulus dimensions, the 4-D and 8-D conditions, respectively. Learning any three dimensions in either condition would lead to 100% accuracy. If people are like these minimal-learning models, they will learn the same number of dimensions in the two conditions (i.e., the minimum necessary) or else will learn the dimensions less well in the 8-D condition.

A related reason to expect reduced learning of individual features in the 8-D case is the *cue competition* (see Kruschke & Johansen, 1999) found in probabilistic cue-learning tasks. Adding features to a category has been shown to reduce the use of individual features as they compete for attention. In probabilistic learning tasks, any particular dimension will be used less in classification when additional dimensions are provided (Edgell et al., 1996), and this effect increases with the salience and cue validity of the additional dimensions (Edgell, Bright, Ng, Noonan, & Ford, 1992; Edgell et al., 1996, Experiment 6). Thus, although adding dimensions to a category might not harm classification of whole items, cue competition should reduce how much each dimension is used and, therefore, the strength of the association between that dimension and the category. Thus, there were a number of reasons to expect that the learning of categories or the categories' individual features should have been harder in the 8-D than in the 4-D condition.

In fact, however, across three experiments, Hoffman and Murphy (2006) found similar learning rates in the 4-D and 8-D conditions. In Experiment 3, the subjects in the 8-D condition actually achieved perfect classification performance in fewer blocks, learned more features (correcting for guessing), and were no slower to categorize those features than were those in the 4-D condition. Thus, Hoffman and Murphy's results appear inconsistent with studies and models suggesting that learning some features of a category should interfere with learning other features.

However, these results do seem consistent with the unlabored speed with which people (and especially children) learn high-dimensional natural categories (see, e.g.,

Bloom, 2000; Carey, 1978; Murphy, 2001). People know dozens of features of everyday concepts, such as cars, cows, and coffee, so it seems unlikely that those concepts were learned by exactly the same mechanism as the one that exhibits blocking, cue competition, and other limitations on how much can be learned. Hoffman and Murphy (2006) suggested that high-dimensional categories may foster learning about the category in general, perhaps by causing learners to allocate free attention to unlearned stimulus dimensions.¹ Rather than treating attention as strictly limited and spread across the stimulus dimensions, they argued that when a category has more dimensions, more attention is recruited to learn about them, if they seem relevant to understanding the category. This may not take place in classical conditioning or probabilistic cue learning, when there is no need to learn about the stimuli beyond giving the right answer in the task.

Prior Knowledge and Feature Co-Occurrence

We can now return to the issue of how prior knowledge might be involved in learning to associate category features to a representation. The goal of understanding feature relations via prior knowledge may, in fact, drive additional learning beyond that which is necessary for classification (Murphy & Allopenna, 1994; Rehder & Ross, 2001). In fact, recent connectionist accounts propose that knowledge encourages learners to relate features either indirectly through previously known concepts or through feature–feature links (Heit & Bott, 2000; Rehder & Murphy, 2003). These mechanisms increase activation to (and therefore, learning of) category labels and features. Perhaps this mechanism would lead to greater feature learning much as does the high-dimensionality manipulation of Hoffman and Murphy (2006).

In fact, there is evidence that prior knowledge can help people learn more about a category. Murphy and Allopenna (1994) observed both faster learning rates (three times faster) and enhanced learning of the category's features when the features were related to one another than when they were not related. In fact, correcting for guessing (see below), the average subject in the knowledge-unrelated condition learned the minimum number of features necessary to achieve perfect classification performance (around 3 dimensions out of 5), but the average subject in the knowledge-related condition learned more (4.2 dimensions). Thus, subjects not only learned faster in the presence of prior knowledge, they learned more about the category even when that additional knowledge was not strictly necessary for accurate performance (see also Kaplan & Murphy, 2000).

Possible Links Between Knowledge Effects and Dimensionality

Apparently, both high-dimensional categories and prior knowledge foster learning of a category's features beyond the minimum necessary for accurate classification. These variables may work via related mechanisms: high dimensionality by recruiting additional capacity for learning interdimensional links, and prior knowledge by using existing links (Rehder & Murphy, 2003). This

suggests that the two variables should interact. Although there is probably some limit on just how many features can be acquired through pure associative learning in a fixed learning phase, this limit can be stretched if features are linked through knowledge. With prior knowledge, learning one feature–category association can activate another, through a knowledge link. In this way, links between the dimensions will feed feature–category associations. For example, if the subject learns that Category A vehicles have the feature *used on glaciers*, the related feature *has treads* might be activated, thereby increasing its association to the same category. When many related features are present, many can take part in this learning process. The difference between how much people learn in high- and low-dimensional concepts should therefore be magnified by the presence of prior knowledge.

The present experiment contrasts learning of lower and higher dimensional categories and also contrasts knowledge-related and -unrelated features. (Because we anticipated that the knowledge-related materials might cause ceiling effects, we used 5 and 10 dimensions, rather than 4 and 8, as in Hoffman & Murphy, 2006.) In each case, a minimum of 3 dimensions was required to learn the categories, and all dimensions were equally diagnostic. We predicted, first, that the knowledge-related categories would lead people to learn more than the minimum number of dimensions. Second, higher dimensional categories should also cause people to learn more properties, even though the task does not require it. The first two predictions follow directly from Murphy and Allopenna (1994) and Hoffman and Murphy (2006), respectively. Our third prediction was that the dimensionality effect would be magnified in the knowledge condition because of the larger number of knowledge links that can be formed in high-dimensional categories. Thus, we predicted that the effect of dimensionality on feature learning would be greater when knowledge was present.

Following the empirical results, we will present simulations comparing Rehder and Murphy's (2003) *knowledge resonance* (KRES) model with a minimal-learning model based on the assumption of independent contributions from stimulus features. The simulations will provide a formal demonstration of how processing feature co-occurrences can lead to enhanced learning of a concept's properties as a function of dimensionality and prior knowledge. We will also confirm that a minimal-learning model cannot account for the dimensionality effect.

EXPERIMENT 1

Method

Subjects. Fifty-six New York University students participated for pay. They were randomly assigned to one of four conditions created by crossing the factors of prior knowledge (knowledge vs. no knowledge) and category dimensionality (5 vs. 10 dimensions).

Materials. The subjects learned categories of vehicles modified from Murphy and Allopenna (1994), using short phrases as features. As Table 1 shows, in the knowledge condition, the features of each category were chosen to be consistent with a theme, whereas in the no-knowledge condition, all the features were generic vehicle properties. In the knowledge condition, the Mobble category contained fea-

Table 1
Features Associated With Each Category (Mobbles and Streaths) for the Knowledge and No-Knowledge Conditions

| Dimension | Mobbles | Streaths |
|--------------|--------------------------|--------------------------|
| No Knowledge | | |
| 1 | One air bag | Two air bags |
| 2 | Colored white | Colored green |
| 3 | Cloth seats | Vinyl seats |
| 4 | Gasoline fuel | Diesel fuel |
| 5 | Rear wheel drive | Front wheel drive |
| 6 | Custom license plate | Generic license plate |
| 7 | Manual transmission | Automatic transmission |
| 8 | Antilock brakes | Normal brakes |
| 9 | Four doors | Two doors |
| 10 | Fast acceleration | High top speed |
| Knowledge | | |
| 1 | Used on glaciers | Used on safaris |
| 2 | Heavily insulated | Lightly insulated |
| 3 | Ice-repellent windshield | Mud-repellent windshield |
| 4 | Colored white | Colored green |
| 5 | Has snow plow | Has vine cutters |
| 6 | Has treads | Has wheels |
| 7 | Heated seats | Air conditioned |
| 8 | Closed roof | Open roof |
| 9 | Made in Norway | Made in Africa |
| 10 | Penguin tracking device | Elephant tracking device |

tures generally associated with a cold weather climate and Streaths with a warm climate. On average, the features in the knowledge and no-knowledge conditions contained 2.4 and 2.3 words, respectively.

The categories were composed of 5 (5-D) or 10 (10-D) binary dimensions, following the structures depicted in Table 2. In the 5-D condition, the subjects learned the category structure shown under 5-D₁ or 5-D₂ in Table 2. As is represented in the table, the Mobble category had 1 as the most common value on each dimension, whereas Streaths had 0 as the most common value on each dimension. However, each dimension had an exception feature, so that it did not perfectly predict category membership.

Table 2 also shows the category structure for the 10-D condition, consisting of Dimensions 1–10. There was one exception feature per item. Thus, both conditions had 10 exemplars per category, but the exemplars in the 10-D condition had twice as many features as those in the 5-D condition. All the dimensions were equally predictive of category membership, in all the conditions.

As Table 2 indicates, we created two versions of the 5-D condition, one using Dimensions 1–5 and the other Dimensions 6–10, so that across the two 5-D conditions all 10 dimensions were shown, allowing comparisons between the 5-D and the 10-D conditions.

Procedure. We presented items randomly in four blocks of 20 trials. One exemplar was presented on each trial, with its features in a random order. The subjects classified the exemplar by pressing the “z” or “/” key on the keyboard. After the response, the correct category label appeared above the feature list, and the word CORRECT or INCORRECT appeared below it for 5.5 sec. This relatively long feedback period was chosen so that the learners in the 10-D conditions would have time to read all the stimulus features. Although current theories do not predict anything specific about the length of feedback influencing the effects of knowledge or dimensionality, it is possible that the longer time allowed the subjects to learn more features than usual. However, the time was constant across conditions and could not be responsible for any of the obtained differences. All the subjects classified vehicles for four blocks.

In the test phase, the subjects viewed one of the learning items or a single-feature item (see below) and classified it as quickly as they could, without feedback. Training and single-feature items were presented twice, resulting in 60 test trials for the 5-D condition and 80 for the 10-D condition.

Table 2
The 5-Dimension (5-D) and 10-D Category Structures

| Stimulus | 10-D | | | | | | | | | |
|-----------------|------------------|----|----|----|----|------------------|----|----|----|-----|
| | 5-D ₁ | | | | | 5-D ₂ | | | | |
| | D1 | D2 | D3 | D4 | D5 | D6 | D7 | D8 | D9 | D10 |
| <i>Mobbles</i> | | | | | | | | | | |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| 4 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 5 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 6 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 7 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 8 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 9 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 10 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| <i>Streaths</i> | | | | | | | | | | |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 |
| 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 |
| 3 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 |
| 4 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 |
| 5 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 |
| 6 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 |
| 7 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 |
| 8 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 9 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 10 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |

Note—The 5-D conditions contained either Dimensions 1–5 or 6–10. The 10-D condition included all 10 dimensions.

To assess the subjects' category knowledge of each individual dimension across conditions, we used items with one feature in *single-feature tests*. However, simply by guessing, the subjects in the 10-D condition would appear to have learned more dimensions, because they have more dimensions on which to guess. We unbiased the estimate of number of dimensions learned by using the guessing correction from Hoffman and Murphy (2006):

$$D_{\text{learned}} = D_{\text{total}} (P_{\text{correct}} - P_{\text{incorrect}}),$$

where D_{total} is the total number of dimensions in the learned category and P_{correct} and $P_{\text{incorrect}}$ are the proportions of correct and incorrect responses to single-feature items, respectively. If half of a subject's guesses are correct, subtracting the incorrect guesses from the correct answers should leave only the known answers (on average). If subjects had solely guessed, they would have as many incorrect as correct answers and would get a score of 0. If they learned all the features, they would receive scores of 5 and 10 in the 5-D and 10-D conditions, respectively. As an alternative measure of individual dimension knowledge, the subjects were recognized as having knowledge of a dimension only if they responded correctly on all four tests of that dimension during transfer. The probability that a subject would meet this criterion on a dimension by chance was only $.50^4 = .0625$. This is a strict criterion, because it does not allow for response errors, but it makes no assumptions regarding guessing. Also, it is a direct measure of whether each individual dimension was learned, rather than an estimate based on overall accuracy.

Results

Learning. We first examined the no-knowledge condition to see whether we replicated the high-dimensional learning advantage found by Hoffman and Murphy (2006, Experiment 3). Note, however, that whereas that study used 4 and 8 dimensions, this one used 5 and 10, so the comparison is not exact. The top of Table 3 shows that 7 (of 14) subjects in the 10-D condition and 5 in the 5-D

condition achieved an errorless block at some point in the learning phase, which suggests a slight learning advantage for the higher dimensional condition. This advantage was supported by a marginally reliable higher proportion correct (a 4% increase) in the 10-D condition, as compared with the 5-D condition [$F(1,52) = 3.42, MS_e = .021, p < .07$]. Although we are cautious in concluding a learning advantage for the 10-D condition, it is clear that the 10-D condition was no harder to learn than the 5-D condition.

We next tested whether we replicated the effect of knowledge on category-learning rate shown in past studies. Overall, the knowledge condition yielded a greater number of perfect performers (16 of 28), as compared with the no-knowledge condition (12). There was also an advantage for the knowledge condition in average proportion correct over blocks (a 3% increase), although this difference was not statistically reliable [$F(1,52) = 2.47, MS_e = .021, p < .13$], most likely due to a ceiling effect, since all the groups were around 90% correct. We found no interaction between knowledge and learning block on proportion correct ($F < 1$).

Whole-item tests. The learning results were similar to the subjects' subsequent classification performance on whole items (unsurprising, given that this test is essentially the same as the learning trials). We found a marginal increase in proportion correct (about 3%, see the middle of Table 3) from the 5-D to the 10-D condition [$F(1,52) = 2.91, MS_e = .004, p < .10$]. The test classifications also yielded the predicted knowledge effect, with knowledgeable subjects performing better (by about 4%) on the subsequent test classifications [$F(1,52) = 7.27, MS_e = .004, p < .01$]. These variables did not interact ($F < 1$). Thus, there is, at best, a weak effect of knowledge on both learning rate and postlearning responses to whole items, but given the apparent presence of ceiling effects, differences are more likely to be found in the single-feature tests.

Single-feature tests. We will focus on number of dimensions learned because proportion correct decreases as the number of dimensions increases when dimension learning is held constant, biasing it against higher dimensionality conditions. The single-feature transfer items

Table 3
Summary of Results, Experiment 1

| Dependent Measure | Dimensionality | | | |
|------------------------------|----------------|------|-----------|------|
| | No Knowledge | | Knowledge | |
| | 5-D | 10-D | 5-D | 10-D |
| <i>Learning</i> | | | | |
| Perfect performers | 5 | 7 | 7 | 9 |
| Proportion correct | .86 | .9 | .9 | .92 |
| <i>Test: Whole Items</i> | | | | |
| Proportion correct | .91 | .94 | .95 | .98 |
| <i>Test: Single Features</i> | | | | |
| Proportion correct | .83 | .72 | .91 | .96 |
| No. learned ^a | 3.25 | 4.36 | 4.07 | 9.29 |
| No. learned ^b | 3.21 | 5.14 | 4.00 | 8.93 |

^aNumber of dimensions learned with the guessing-corrected estimate. ^bNumber of dimensions learned with the alternative criterion.

also will allow us to compare response times (RTs) across conditions, which is not possible with the whole-item or learning analyses, because 10-D items would always take longer to read than the 5-D items. Here, RTs to the same stimuli will be compared.

The second to last row of Table 3 lists the number of dimensions learned in each condition (guessing corrected). The first two columns in the table show that in the absence of knowledge, increasing dimensionality yielded an increase in the number of dimensions learned, from about three dimensions to four ($SDs = 1.30$ and 2.30 , respectively). This increase was consistent with the effect of dimensionality observed by Hoffman and Murphy (2006), in which people learned more than the minimum number of dimensions in the high-dimensional condition.

In the present study, we predicted that knowledge relations would feed learning of feature–category associations, causing an even larger effect of dimensionality in the knowledge condition. We found this effect, shown in the last two columns of Table 3. When the features were thematically related, increasing dimensionality increased the number of dimensions learned from about four dimensions to nine ($SDs = 1.30$ and 1.09 , respectively). This increase in the number of dimensions was about five times larger than the corresponding one in the no-knowledge condition. In short, dimensionality strongly interacted with knowledge.

To test the reliability of this interaction we compared the number of dimensions learned across the four conditions in a 2×2 between-subjects ANOVA. The interaction between knowledge and dimensionality was, in fact, highly reliable [$F(1,52) = 24.0$, $MS_e = 2.459$, $p < .001$]. The interaction suggests that the effect of knowledge is to magnify the effect of dimensionality. Whereas the advantage for the 10-D condition was marginally reliable in the no-knowledge condition [$F(1,52) = 3.44$, $p < .07$], it was highly significant in the knowledge condition [$F(1,52) = 57.37$, $p < .001$]. The advantage seen in the 10-D no-knowledge condition replicated a similar comparison in Hoffman and Murphy (2006), albeit marginally (but see the next paragraph). However, the effect size of dimensionality was about 10 times larger in the knowledge condition ($\eta_p^2 = .59$) than in the no-knowledge condition ($\eta_p^2 = .06$).

These results are supported by the alternative measure of feature knowledge requiring perfect performance on a dimension to indicate learning (see the bottom row of Table 3). The contrast between 5-D and 10-D conditions was reliable both with knowledge [$F(1,52) = 70.03$, $MS_e = 2.40$, $p < .001$] and, now, in the no-knowledge condition [$F(1,52) = 10.53$, $MS_e = 2.40$, $p < .01$], replicating the dimensionality effect. The predicted interaction was also present here [$F(1,52) = 13.10$, $MS_e = 2.40$, $p < .001$]. The effect of dimensionality was about four times greater in the knowledge condition ($\eta_p^2 = .63$) than in the no-knowledge condition ($\eta_p^2 = .17$), using the alternative measure.

Truncation explanation. One concern is whether the increase in dimensions learned in the 10-D conditions was due to a limit on the number of features one can learn in

the 5-D condition. That is, the 5-D subjects could not score higher than five, whereas the 10-D subjects could. We can address this concern by comparing the number of subjects in the 5-D condition who learned all five dimensions with those in the 10-D condition who learned five or more: The 10-D subjects no longer get extra credit for learning more than five dimensions in this comparison. There were 2 subjects in the no-knowledge, 5-D condition who learned all five dimensions, as compared with 6 (basic criterion) or 8 (alternate criterion) in the corresponding 10-D condition. Because fewer subjects learned five or more dimensions in the 5-D than in the 10-D condition, the former's lower learning cannot be explained by truncation.

Reflecting the magnification effect of knowledge on dimensionality, there is a pronounced difference between the number of subjects learning five or more dimensions in the 10-D and 5-D knowledge conditions (14 vs. 7). Regardless of criterion, *every* subject in the knowledge 10-D condition learned five or more dimensions. Thus, none of the obtained differences between the 5-D and the 10-D structures can be explained by truncation in the 5-D condition, since more subjects learned at least five dimensions in the 10-D condition.

Reaction times. We examined RTs to determine whether there were speed–accuracy trade-offs in the subjects' single-feature responses. In fact, RTs were very similar between the 5-D ($M = 1,413$, $SD = 478$) and the 10-D ($M = 1,422$, $SD = 302$) conditions ($F < 1$), and the subjects in the knowledge condition ($M = 1,291$, $SD = 354$) were 253 msec faster than those in the no-knowledge condition ($M = 1,544$, $SD = 402$) [$F(1,52) = 8.23$, $MS_e = 0.047$, $p < .01$]. There was no interaction between knowledge and dimensionality in RTs [$F(1,52) = 2.00$, $MS_e = 0.047$, $p < .17$].

The shorter RTs in the knowledge condition are not a readily interpretable effect, because comparison of different features is involved, but the result nevertheless rules out the possibility of speed–accuracy trade-offs. The more accurate conditions were responded to at the same rate as or more quickly than were the less accurate conditions in every comparison.

EXPERIMENT 2

A key result from Experiment 1 is the interaction of knowledge and number of dimensions for the number of features learned. Expanding the number of dimensions of the stimuli caused people to learn more in the no-knowledge condition (as in Hoffman & Murphy, 2006), but the effect was relatively small. In the knowledge condition, the number of features learned essentially doubled when the number of dimensions doubled. We have been interpreting this as people's actually learning more of the dimensions that are presented, through the knowledge connections that permeate the thematically related features. That is, learning that Streaths have vine cutters helps in learning that they are made in Africa.

However, our results are compatible with an effect of knowledge at retrieval, rather than—or in addition to—initial learning. (We thank a reviewer for raising this possibility.) For example, perhaps people did not, in fact, learn

that Streaths were made in Africa but, when given the test at the end of the experiment, they inferred this fact, using the features that they *had* learned. This possibility seems most likely in the 10-D knowledge condition. In the no-knowledge conditions, people could not infer unknown features from known ones, since they were not related. In the 5-D knowledge condition, the subjects learned only a little more than did those in the no-knowledge condition; thus, inference could not have had much effect there. But in the 10-D condition, learners might have truly learned only four or five dimensions and then inferred the categorizations of the rest, thereby leading to a large figure for number of dimensions “learned.”

As will be seen in the modeling we report below, effects of knowledge during test are completely compatible with theories of how knowledge influences category learning. To foreshadow, KRES assumes that features related to a theme are linked in a way that tends to increase the activation going to the correct category at test. So, effects at test are by no means incompatible with such accounts. Indeed, we have no a priori way of knowing how much of a knowledge effect takes place at learning and how much at test. However, to address our goal of comparing how much is learned about categories’ features, we set out to discover whether the subjects’ performance in the 10-D knowledge condition could be explained by their inferring, rather than learning, more features than did those in the 5-D condition.

In Experiment 2, we taught the subjects 5-D knowledge-related concepts but then tested them on all 10 knowledge-related dimensions. In this way, we were able to directly compare people’s accuracy and speed in classifying features that had probably been learned with their accuracy and speed for those that must have been inferred. Our expectation was that people would be less accurate and slower to identify inferred features, which would allow us to rule out the possibility that most of the knowledge-related features in Experiment 1 were inferred. Because people would likely be slower to read and comprehend features that they had never seen before, we had a preexposure phase in which the subjects read the features in a different context, to control for the effect of mere exposure on the learned versus inferred features.

Method

Twenty New York University students participated for pay. The subjects were randomly assigned to one of the two 5-D knowledge conditions from Experiment 1 (either side of Table 2). The stimuli and procedure were identical to those in the 5-D knowledge conditions in Experiment 1, with the following differences.

Inference tests. After the learning phase, we assessed the subjects’ category knowledge through single-feature tests, as in Experiment 1. However, in addition to classifying the *old features*, the subjects were asked to infer the category membership of features from the opposite 5-D knowledge condition, so that each subject was tested on all 10 knowledge-related dimensions. The subjects were not warned that they would be tested on new features.

Preexposure task. Experiment 2 was intended to contrast RTs and accuracies between old and inferred features from the single-feature tests. One complication, however, is that the old features, but not the inferred features, would have been seen during the classification-learning phase of the experiment. Longer RTs for the

inferred features might reflect not just the inference process, but also the time required to read and comprehend new features.

To minimize reading differences, we preexposed subjects to the inferred features before the classification-learning phase, using a memory task that would not integrate those features into the categories used in the classification phase. The preexposure task included 20 study-and-test trials, in which the subjects viewed a study screen with a randomly ordered list of six features, numbered 1–6. Two features were inferred feature items, and four were filler items chosen so that the resulting list would not be interpreted as describing a vehicle (see the Appendix). After studying the list at their own pace, the subjects saw a second list in which one randomly selected filler feature had changed (e.g., *can take photographs* was changed to *can record sound*). The subjects then indicated which of the six features had changed by pressing the corresponding number key. Filler features and items were selected randomly on each trial without replacement, so that by the end of the preexposure phase, each inferred feature had been read a total of eight times (i.e., at study and at test, on four study–test pairings).

Results

The subjects performed well in the preexposure phase, at 85% accuracy (chance accuracy was about 17%), showing that they had attended to the presented features. Thus, we were certain that they had read and encoded the features that would be used as inferred features in the test phase.

Following preexposure, the subjects learned the 5-D knowledge-related categories and were tested on both learned and inferred single features. The accuracy measure that distinguished those features best was the strict criterion metric of number of features learned (requiring perfect performance on that dimension). Because this measure requires good performance right from the beginning, it is more useful for distinguishing learned from inferred features than is a measure that could be diluted by further learning during the test phase. If the subjects were inferring features, they might make mistakes when they first encountered a new feature but begin to make the connection between that feature and the themes underlying the categories on a second or third exposure. According to the strict criterion, the subjects performed nearly perfectly on the 5 old dimensions, learning 4.7 of them. The subjects responded perfectly to an average of 3.8 of the inferred dimensions. It is clear that people were, in fact, able to infer the category membership of a majority of the new features. However, they did not respond as accurately to the inferred dimensions as to the trained dimensions [$t(19) = 3.32, p < .01$]. On the basis of the learners’ high accuracy on the inferred features in this experiment, the concern over whether the observed differences in Experiment 1 were based on inference seems valid.

We next compared the classification RTs in the first block of the test phase, avoiding any influence of repeated classifications of the new features in the second block. As we expected, people were reliably faster to classify the old features ($M = 1,207, SD = 218$) than the new ones ($M = 1,504, SD = 417$) [$t(19) = 3.28, p < .01$]. Interestingly, this ~300-msec difference completely disappeared in the second block of testing, with average RTs of 1,042 msec ($SD = 208$) and 1,041 msec ($SD = 210$) for the old and new items, respectively. This demonstrates how effortlessly knowledge-based features can be learned: After only

a single exposure without feedback in the first test block, they became as strongly associated to the category as did features that had been viewed many times with feedback. But importantly, this difference also contradicts the concern that people were inferring correct responses to many of the knowledge-based features in Experiment 1. That is, the speedup after the single exposure shows that the subjects were not inferring the features in later test blocks but, rather, remembered them and that very little exposure, and no feedback, is necessary to acquire the new feature. It would be very peculiar if the subjects in Experiment 1, with explicit learning trials, did not learn the knowledge-based features that the subjects in Experiment 2 learned without feedback.

Reanalysis of Experiment 1. We cannot directly compare these RTs with those in the first experiment, given the differences in procedure (the initial task, slightly different test instructions) and the absence of random subject assignment to the two experiments. However, we can use the results obtained in this experiment to direct a reanalysis of the earlier results to see whether they show signs of feature inference.

As was discussed earlier, it is the 10-D knowledge case that would be most subject to any effect of inference, because it had the greatest number of features learned by far. In contrast, the knowledge and no-knowledge groups learned about the same amount in the 5-D categories, so there is no reason to suspect that inference unduly aided the knowledge group there. (Keep in mind that inference works only after one has learned some of the features, which then allow extraction of the theme. So, even in the knowledge condition, some features must be actually learned before others can be inferred.) Therefore, the critical question is whether the interaction of knowledge and dimensionality resulted primarily from the fact that people in the 10-D knowledge condition used inference at test.

To examine this, we compared the classification RTs of the 5-D and 10-D knowledge conditions in the first test block of Experiment 1, since that is the block that showed a difference between inferred and learned features in Experiment 2. If people were inferring many of the features in the 10-D knowledge condition, but not in the 5-D knowledge condition, we should observe an RT advantage in the latter. In fact, the pattern was in the reverse direction, with faster responses in the 10-D knowledge condition ($M = 1,359$ msec, $SD = 312$) than in the 5-D condition ($M = 1,594$ msec, $SD = 580$). To identify the effects of knowledge and dimensionality, we performed a two-way ANOVA on the first test block RTs and discovered that knowledge features were classified 331 msec faster, on average [$F(1,52) = 6.22, p < .05$]; again, this is not readily interpretable, given the different features in the two conditions. There was also an interaction, resulting from the fact that higher dimensionality increased speed for knowledge-related features but decreased it for neutral features [$F(1,52) = 3.94, p = .05$]. This pattern is the opposite of the one that would have been expected if the 10-D knowledge condition had consisted primarily of inferences. Given that inferring is slower than classify-

ing learned features (from Experiment 2), that condition should have been the slowest—not the fastest. We can therefore conclude that the apparently learned features in Experiment 1 were, in fact, likely learned at training and were unlikely to have been mostly inferred at test time. Of course, this does not deny that inferring does occur in knowledge-based categories; it no doubt often does. But in the present experiment, learning the features was apparently so efficient that inferring was not necessary.

This RT interaction may also reveal the nature of the processing differences for knowledge-based and neutral features. When one learns more dimensions without knowledge, the RT goes up slightly, presumably because of cue competition effects, but also because people may not have learned all the dimensions and so are spending more time in the first test block trying to recall the answer. In contrast, when the features are thematically related, they gang up and aid each other's processing. That is, knowing that Streaths are made in Africa helps one learn that they have vine cutters. The two features reinforce one another, thereby increasing their association to the category, rather than competing with one another. In the next section, we will discuss one possible way of instantiating these principles in a computational model.

The interaction between number of dimensions and prior knowledge is related to findings in the recognition memory literature. Normally, the more facts one studies about a single item, the slower the retrieval is of any of those facts in a recognition test. However, when the facts are thematically related, this effect disappears (Smith, Adams, & Schorr, 1978). Reder and Ross (1983) showed that this could be explained by people's judging the thematic consistency of thematically related facts, a strategy that is impossible for unrelated facts. When this strategy was prevented, interference among the facts reappeared. Interestingly, people's explicit judgments of thematic consistency became faster and more accurate the greater the number of thematic facts they had learned (Reder & Ross, 1983, p. 63)—analogous to our RT findings. Because our subjects were asked to classify each feature—not to recognize it—our task could be considered as a kind of thematic consistency judgment (assuming that people learned the themes associated with each category). But the results from our Experiment 2 show that even thematic consistency may be easier to judge when the feature has already been learned. Part of learning the theme underlying our categories involves learning how the features relate to the theme, and this requires specific encoding of the features, rather than simply representing the gist of the category. Nonetheless, the convergence of our results and those in the older recognition memory literature is intriguing.

MODEL-BASED ANALYSES

The psychology of concepts contains many models that describe how people learn categories of a few simple dimensions. Such models do well in characterizing people's behavior in learning simple concepts when there is little or no connection to prior knowledge. The question is, what

is needed to explain everyday concept learning, where concepts are composed of many dimensions and where concepts are not merely arbitrary collections of features?

This question motivated the present investigation of the combined influence of dimensionality and prior knowledge on category learning. According to minimal-learning models, people should stop learning feature–category associations once they can reliably predict the category. When learning is faster, as in the present knowledge or dimensionality effects, there is less opportunity to learn nonminimal properties of the category. If the subjects in our task were attempting to learn simple rules, the same number of dimensions should have been learned across dimensionality conditions (because all our categories can be learned by any three dimensions). Alternatively, if the subjects in our task were attempting to learn feature–category associations, higher dimensionality should yield reduced single-feature learning, due to increased cue competition. Thus, minimal-learning theories predict either that the same number of features should be learned or that learning of the features will be weaker in high-dimensional conditions, as reflected in lower accuracy or longer RTs on single-feature tests. However, in Experiment 1, single-feature tests showed that the subjects in the 10-D condition learned more dimensions than those in the 5-D condition did—a number greater than the three required for perfect performance—yet RTs were equal across conditions. Feature learning was thus stronger than would be predicted by minimal-learning models.

Enhanced feature learning from dimensionality and knowledge has been observed now in multiple studies, including the present Experiment 1, across different category structures and over a range of stimuli (drawings of bugs and textual descriptions of vehicles). New here is the interaction, where prior knowledge magnified the dimensionality effect, suggesting that the two variables influence a common learning mechanism. Having less error did not seem to harm single-feature learning. In fact, these results provided a rare (if not unique) demonstration of cue cooperation, rather than competition. How could the subjects have learned more in the high-dimensional and knowledge conditions?

One proposal was that both knowledge and dimensionality increase categorizers' attention to feature co-occurrence. If knowledge is represented (in part) as interfeatural relationships, learning one feature–category association can lead to learning others that are connected through the same network of knowledge structures. Thus, knowledge can lead to a greater amount of activation of (i.e., attention to) feature–category associations and, thereby, enhance learning. The attention hypothesis also explains the interaction between knowledge and dimensionality: High dimensionality increases the knowledge effect, because, with many dimensions, there is a larger number of interfeatural relationships that feed the learning of feature–category associations.

As the number of dimensions increases in the absence of knowledge, however, attention shifts to feature co-occurrence in a different way: Subjects are motivated to

search among the many dimensions for explanations of feature co-occurrence. When knowledge structures are not available to facilitate learning, subjects may engage in a process of inventing them for future use. For example, when learning categories of vehicles in the 10-D, no-knowledge condition, subjects may look for an explanation for why rear wheel drive vehicles also typically have a manual transmission. Perhaps manual transmission is easier to manufacture with rear wheel drive.

The attentional strategy hypothesis makes sense if categories are not random collections of features. People may expect a reason for features' co-occurrence (Medin & Ortony, 1989) and, moreover, may expect that it is useful to know the reason. For example, consider the novice mechanic distinguishing old and new engines. Although the model year alone suffices, the mechanic can learn that older engines use a carburetor and newer engines use fuel injection. These features enable him or her to diagnose and then repair engine problems associated with one or the other fuel system. Whereas relating engine age and fuel system helps produce a useful and flexible concept, learning to distinguish engines with the minimum number of features (e.g., model year alone) does not (Markman & Ross, 2003). Thus, adding dimensions to a category may encourage categorizers to learn about relations between dimensions that are not necessary to learn a *particular* distinction but are likely to be useful for learning a broad range of conceptual distinctions (e.g., electrical vs. hydrogen engines), inferences (e.g., fuel efficiency), or purposes (e.g., diagnostics, repair, or innovation).

This attentional hypothesis is one of several potential explanations that share a common thread—a nontrivial relationship between co-occurrences of features. The following section will consider the implications of an alternative hypothesis that assumes no relationship between features. Instead, each feature has an association with a category label, and each of the perceived features *independently* contributes to the category response.

Independent Feature Models

There are a number of models that make the assumption that features are independent sources of information about categories. In the perception literature, models that explicitly implement this independence assumption have been quite prominent (Movellan & McClelland, 2001). In these models, the presence of each feature contributes weight toward a potential response. The weights of all features are added, and the result is transformed into a response probability.

One such model, the fuzzy logical model of perception (FLMP; Oden & Massaro, 1978), assumes that stimulus features and category responses are linked by an association ranging in strength from 0 (*perfect negative association*) to .5 (*no association*) to 1 (*perfect positive association*). Since each association strength is assumed to be independent, this formulation allows for an easy representation of partial stimuli. Massaro and Friedman (1990) have provided an equation that relates the response probabilities of a whole item to the response probabilities of its two features. Here, we want the re-

verse, the response probability for single features as a function of whole-item response probabilities. By generalizing Massaro and Friedman's equation to multiple features and inverting it, we get the following:

$$E(SF) = \frac{\left(\frac{W}{1-W}\right)^{\frac{1}{D_{\text{total}}}}}{1 + \left(\frac{W}{1-W}\right)^{\frac{1}{D_{\text{total}}}}}, \quad (1)$$

where $E(SF)$ is the expected single-feature accuracy, W is whole-item response accuracy, and D_{total} is the number of dimensions.²

Another independent feature model is the canonical minimal-learning model, the Perceptron (e.g., Cohen & Massaro, 1992). The weights of the Perceptron convert input activations to output activations and can be trained through error-correcting feedback. Although the Perceptron is not considered a modern model of classification learning, it is closely related to the classic Rescorla-Wagner learning theory. We consider the consequences of assuming an independent feature model by simulating how such a model would perform in our experimental task, to demonstrate that the observed effects require non-independence of features.

We implemented a Perceptron model with a single output node whose nonlinear activation ranged from 0 (the Mobble category) to 1 (the Streath category). There were either 5 or 10 input nodes, whose activations were set to 0 or 1, following Table 2. We used an entropy-based error function and weight update rule that allows the output activation to be treated as a response probability (Hertz, Krogh, & Palmer, 1991; Hopfield, 1987). (To allow output activations to represent response probabilities, we used a continuous sigmoid activation function, rather than the hard threshold of the original Perceptron.)

Whole items and single features. We trained the Perceptron on the 5-D and 10-D category structures (without

knowledge), using four blocks of training that were the same as those the subjects received. Model parameters (the learning rate, the range of initial weights, and α , the slope parameter of the activation function) were tuned to match the subjects' average whole-item and single-feature results, using a chi-square goodness-of-fit measure. In addition, the model's predictions over a wide range of parameter settings were collected to allow a qualitative measure of performance. We first examined whether the Perceptron can be tuned to account for the subjects' single-feature performance for categories without prior knowledge. Table 4 shows that the Perceptron was, in fact, off target, with the best fits typically severely underestimating single-feature accuracy and number of dimensions learned for both 5-D and 10-D categories. (The best-fitting results shown in Table 4 were for the following parameters: learning rate = 1.2, $\alpha = 0.1$, initial weight range = 0.1.)

We were additionally able to show that this failure of the Perceptron reflects a general property of the independent feature assumption. Using the FLMP predictions of Equation 1, we calculated the expected single-feature accuracy, $E(SF)$, from the empirically observed whole-item accuracy (i.e., inserting the W values from the table into Equation 1 for each row). Single-feature accuracy values near the expected value indicate independent dimensional combination, but values higher than expected indicate that dimensions are not being treated as independent sources of information about the category label. The third and fourth columns of Table 4 reveal that single-feature accuracy was higher than was theoretically predicted. This result reflects the fact that people learned more dimensions than were expected from their whole-item accuracy, assuming an independent combination of dimensions.

Although the FLMP does not well describe human performance here, its predictions are, in fact, essentially identical to the Perceptron's SF and whole-item response proportions, over its parameter space. Figure 1 (top row) shows the range of the Perceptron's response patterns (black dots) and the average subject data (large gray dot),

Table 4
Model-Fitting Results

| Model | D_{total} | W | SF | $E(SF)$ | No. Learned ^a | No. Learned ^b | $E(\text{Learned})$ |
|--------------|--------------------|-----|------|---------|--------------------------|--------------------------|---------------------|
| No Knowledge | | | | | | | |
| Empirical | 5 | .91 | .83 | .65 | 3.3 | 3.2 | 2.4 |
| | 10 | .94 | .72 | .59 | 4.4 | 5.1 | 2.7 |
| KRES | 5 | .92 | .80 | .66 | 3.0 | 2.4 | 2.1 |
| | 10 | .96 | .74 | .60 | 4.8 | 5.1 | 3.0 |
| Perceptron | 5 | .90 | .64 | .64 | 1.5 | 0.9 | 0.9 |
| | 10 | .96 | .60 | .60 | 1.9 | 1.3 | 1.3 |
| Knowledge | | | | | | | |
| Empirical | 5 | .95 | .91 | .69 | 4.1 | 4.0 | 3.4 |
| | 10 | .98 | .96 | .62 | 9.2 | 8.9 | 8.5 |
| KRES | 5 | .96 | .95 | .70 | 4.5 | 4.1 | 4.1 |
| | 10 | .98 | .97 | .62 | 9.4 | 8.8 | 8.8 |

Note— D_{total} is the number of dimensions. W is whole-item accuracy. SF is single-feature accuracy. $E(SF)$ is the expected single-feature accuracy based on the whole-item accuracy and the assumption of independent contributions to responses. $E(\text{Learned})$ is the expected number of single-features learned according to the strict criterion, assuming equal attention to features. ^aGuessing-corrected estimate. ^bStrict criterion.

where each location in the space represents a particular combination of whole-item and single-feature accuracy. The solid black line is the theoretical curve corresponding to an independent, additive combination of dimensions according to the FLMP. The Perceptron's performance closely follows the FLMP curve, which is not near the subjects' data. Thus, the Perceptron failed here exactly because it assumes that information from separate dimensions is combined independently.

Uniform learning of features. We observed in the analysis above that the Perceptron independently combines dimensions. We now will describe another interesting property of the Perceptron, which can serve as a second point of comparison with human learning. Consider the following question: Did people learn all dimensions uniformly, or did they learn a subset of dimensions particularly well? Consider two ways for a subject in the 10-D condition to score 75% correct on single-feature tests. One way is to answer correctly on three of the four

single-feature tests for each dimension; the subject would thus demonstrate moderate and uniform learning of every dimension. But a subject can also get 75% by answering correctly on all four single-feature tests for five of the dimensions but answering correctly on only half of the single-feature tests—guessing—for the remaining five dimensions. In this case, learning is nonuniform; five dimensions were answered perfectly, and the others were answered at chance levels.

With uniform learning, the expected number of dimensions learned with our strict alternative criterion (all four single-feature responses for a dimension correct) is the following: $E(\text{learned}) = D_{\text{total}} * SF^4$ —that is, the single-feature accuracy, raised to the fourth power to give the likelihood of four correct responses for a particular dimension, times the total number of dimensions. If the number of dimensions responded to perfectly is higher than expected, subjects must have learned some dimensions particularly well (and others less well). The expected

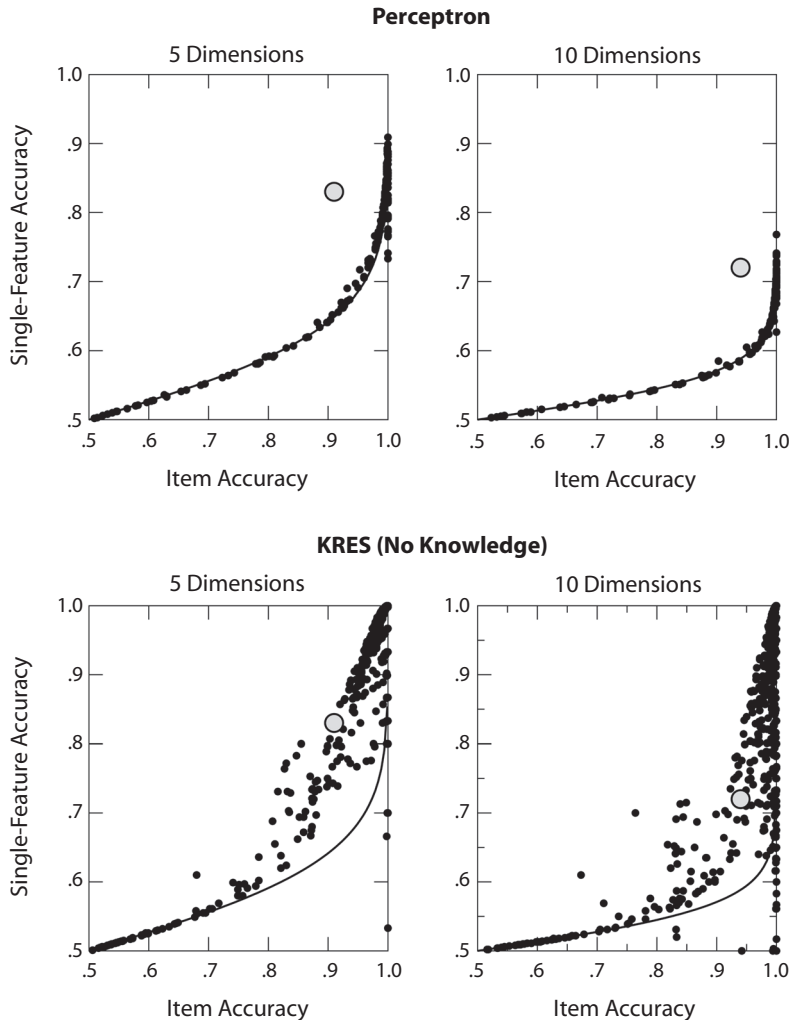


Figure 1. Range of possible model predictions of Perceptron and the knowledge resonance model (KRES; black dots), with the empirical results in the no-knowledge case (large circle) and the relationship between whole-item and single-feature accuracy as predicted by the fuzzy logical model of perception (solid line).

and actual numbers of dimensions learned appear in the second to last and last columns of Table 4, respectively. In the no-knowledge condition, the number of dimensions learned was higher than expected, especially in the 10-D case. Clearly, the subjects in no-knowledge conditions learned some dimensions better than they did others. However, this difference disappeared in the knowledge conditions (in part, because learning was so high there).

Examining these last two columns in Table 4 for the Perceptron reveals that, unlike humans, the model learned the dimensions uniformly. Further analysis indicated that the Perceptron showed nonuniform learning only in circumstances in which the learning rate parameter was pathologically large and the weights saturated in only one or two learning steps. Otherwise, the gradient descent learning rule evens out any initial variance in the weights as it minimizes error, resulting in all dimensions being learned equally well.

In summary, the FLMP and Perceptron, which represent independent learning of each stimulus dimension, cannot account for the results. Unlike these simple models, people learned more dimensions than would be expected, and their associations were clumped, with some dimensions learned well and others not learned. Humans do not appear to assume independent features when they are learning about novel categories. This finding may provide a useful constraint on a number of models of category learning.

The Knowledge Resonance Model

The modeling above highlights two reasons why independent feature models such as the Perceptron are inadequate to account for our empirical results.³ To account for our results, we need a model with at least three properties. First, its responses to single features and whole items should not reflect the independent feature assumption. Second, it should not always learn features in a uniform manner. Any model based on the FLMP or on the Perceptron will violate these two requirements.

The third property of a successful category-learning model is that it should account for the effects of prior knowledge, which few current models can claim. However, it is possible to account for prior knowledge effects while still violating the first two requirements of nonindependence and nonuniformity. For example, one model of category learning with prior knowledge is the Baywatch model (Heit & Bott, 2000). Baywatch adds prior knowledge nodes to a single-layer neural network model of categorization. Although Baywatch can account for many effects of prior knowledge on category learning, it reduces to a Perceptron when knowledge is absent and, as a result, necessarily fails to account for the nonindependence and nonuniformity requirements. Therefore, we explored the performance of Rehder and Murphy's (2003) KRES model on our task.

We chose KRES because it can incorporate prior knowledge and is the only such model with interactive activation processes. KRES was designed to account for many knowledge effects, including accelerated category learning, better learning of features not related to prior knowledge, and reinterpretation of features in light of error-corrective feedback. The present simulation, however,

will test whether KRES generalizes to a phenomenon for which it was not originally designed. As we will describe in detail below, KRES's mechanisms for learning and representing knowledge should satisfy the three requirements of increased feature learning relative to whole-item accuracy, nonuniform feature learning, and the interaction between number of dimensions and prior knowledge. Furthermore, its interactive architecture may lead to effects of feature co-occurrence in higher dimensional categories, as revealed in the present study. At present, however, KRES lacks an attentional-learning mechanism, so any success it has in modeling the task will not be due to explicit attention differences across learning conditions, as has been proposed by Hoffman and Murphy (2006). Instead, such success would be due to the model's interactive activation processes. Thus, KRES tests our assumptions about how knowledge may influence feature learning and may provide another account of the dimensionality effect.

Figure 2 illustrates the KRES model, as configured for the 5-D knowledge condition. Like the Perceptron, KRES is a prototype model, with weighted connections between input feature nodes and category label output nodes. Each input and output dimension is represented by two mutually exclusive feature nodes. KRES's output activations can be transformed by the standard Luce choice rule to yield category response probabilities. KRES can represent knowledge either as connections to prior knowledge nodes (as in Heit & Bott, 2000), or as connections between input nodes, indicating prior associations among features. (Real knowledge is undoubtedly more complex; the feature links are a simplification that stand in for causal and other knowledge that connects properties of a concept.) Because the vehicles learned in our experiment did not refer to previously known categories, the prior knowledge nodes were not used, and knowledge was represented only as links between input features.

In comparison with other models, KRES has two important properties. First, activation in the model resonates among bidirectional links in a type of constraint satisfaction process. This means that representations in the input layer can change substantially after the initial stimulus presentation, once the model takes other constraints into account (i.e., once activation flows to and from other nodes in the network and the network settles). This property may increase single-feature performance, especially in the high-dimensional condition, as activation flows downward from the activated output nodes to consistent input features, then back up again to the output nodes. Note that in both the attention hypothesis and KRES, the linking of features to each other, directly or indirectly, plays a central role.

This interactive activation process is the basis for the biologically plausible contrastive Hebbian learning (CHL) rule (O'Reilly, 1996), which updates weights in a KRES network based on feedback (Rehder & Murphy, 2003). Briefly, the input to the network is presented, and activation is allowed to flow throughout the network until it settles into a stable state. The activation of the output nodes is transformed to a response prediction. Then the target value of the output nodes is added to those nodes,

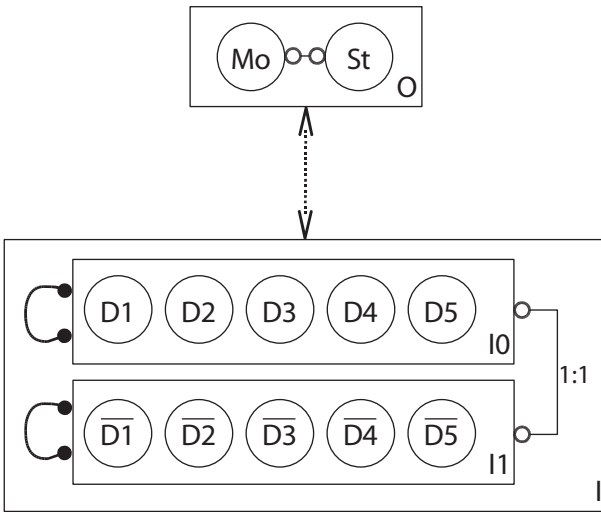


Figure 2. The 5-D KRES model as used in simulations, with prior knowledge connections shown. The 10-D version uses 10 input nodes, whereas the version with no prior knowledge does not have the excitatory lateral connections. Excitatory connections are shown with solid circles, and inhibitory connections are shown with empty circles. I, input nodes, and I0 and I1 represent the nodes associated to the two themes. Each dimension in I0 has an inhibitory link to its opposite node in I1. O, output nodes, the two categories Mobbles and Streaths.

and activation again flows throughout the network until it again settles. A Hebbian process then uses the differences in each node's activation between the two settled states to update the weights between each pair of nodes.

KRES's second notable property is that it can represent prior knowledge as direct lateral connections among related features. Activation of one input feature spreads to directly linked features. These links boost node activation, with two potential effects on the model's performance. First, the links can combine to facilitate classification learning. As the CHL learning rule strengthens weights in proportion to the activation levels of linked nodes, the activation from the knowledge will result in stronger associations being learned. The second effect of the lateral links is to increase single-feature classification accuracy. At test time, activation will spread from feature to feature, through the lateral connections and then to output units, increasing output node activation and, therefore, single-feature performance. Thus, KRES's account of knowledge has similarities to Hoffman and Murphy's (2006) attention strategy hypothesis, in that knowledge causes more features to become activated and weights to be more strongly learned.

KRES was trained on 5-D and 10-D category structures. The feature values of 0 and 1 were converted into pairs of inputs (+1, -1 for a 1; -1, +1 for 0), since KRES represents each dimension as two separate units (one for each value). For single-feature tests, the missing dimensions were set to (0, 0). We simulated knowledge and no-knowledge conditions by configuring KRES with and without lateral links between input nodes, respectively. The same parameters were used for all four KRES configurations, simulating the four experimental conditions. KRES has five tunable parameters: learning rate, an α parameter representing the

slope of the activation function, the weights of the fixed inhibitory and fixed excitatory weights, and the initial range of trainable weights. Parameters were tuned via a grid search over the parameter space.

Simulation results. Table 4 shows that, unlike the independent feature Perceptron model, KRES could account for the relatively high single-feature accuracy and number of dimensions learned. KRES's resonance properties allowed it to take advantage of feature co-occurrence and to both qualitatively and quantitatively account for subjects' single-feature accuracy performance. (The results in Table 4 were found with learning rate = 0.25, α = 0.75, inhibitory fixed weights = -2.5, excitatory fixed weights = 0.5, and initial trainable weights = 1.0.)

We next examined whether, by adding lateral knowledge links among input nodes, KRES could account for the knowledge effect and its interaction with dimensionality. Recall that no KRES parameters were changed in modeling the knowledge data. Nevertheless, Table 4 shows that KRES was able to account for the enhanced whole-item and single-feature performance in the knowledge condition.

Given that KRES was designed to account for the learning-enhancing effects of knowledge, it was perhaps not too surprising that it was able to account for the main effect of knowledge found in our data. It is more impressive that the model also accounts for the dimensionality effect and its interaction with knowledge. Table 4 shows that KRES yielded a much larger dimensionality effect in the knowledge condition, as was observed in the human data.

KRES was able to account for all observed effects in our initial model analysis. However, it is possible that it accounted for the effects through its functional complexity, rather than any inherent explanatory property. In other words, it may not be that KRES, in general, predicts the observed pattern of results, but rather that it found an unlikely region in its parameter space that incidentally matched the data (Pitt, Kim, Navarro, & Myung, 2006). However, Figure 1 (bottom row) shows that this is not the case. In contrast to the Perceptron, KRES patterned near the human subject data and away from the FLMP curve over a wide sampling of its parameter space. KRES patterns this way because, when making predictions based on single features, activity resonates from the single presented feature to the category node and back to input nodes, changing their activation to be consistent with the category prototype and, thereby, yielding greater activation of the category node. This interactive flow of activation in single-feature tests is even stronger with knowledge, because associated input nodes are connected to each other directly. (However, when all dimensions are present in the stimulus, input node activation is less affected by interactive feedback, and category activation does not change as much.) Given that KRES and the Perceptron have a similar feature-to-category node architecture, we can safely conclude that it was KRES's interactive activation properties, and not any superficial functional complexity, that allowed it to match the observed pattern of behavior.

As is also shown in Table 4, KRES was able to capture the nonuniform dimension-learning effect. This result has

to do with an aspect of KRES's learning rule. Instead of the target value *replacing* the output unit's activation, as in the Perceptron learning rule, in CHL, the target value is *added* to the output unit's activation. As a result, any initial variance in weights is often preserved, allowing nonuniform response patterns even after extensive learning. The initial variance in weights between inputs and outputs might correspond to subjects' initial guesses as to which features are likely to be important in the learning task. Although this is not hypothesis testing in the traditional sense, subjects may, in fact, begin the experiment with hypotheses about which subset of dimensions is likely to be most predictive of the categories. This initial guess causes differences in association weights, leading to nonuniform learning.

The combination of CHL and input-output weight variance is not the only way to capture nonuniform learning in KRES. In the current simulations, we assumed, for simplicity, that all of the learners in the no-knowledge condition had zero associations between features. But this might not be entirely accurate. For example, there may be a weak association between rear wheel drive and manual transmission, whereas there may be no association at all between gasoline fuel and cloth seats. Thus, KRES has a second way of accounting for the nonuniformity of learning in the no-knowledge condition. Regardless of the potential source of variation, it is clear that KRES captures the nonuniform learning of dimensions that people exhibit.

It is also worth noting that KRES was able to capture learners' more uniform learning in the knowledge condition, due to KRES's lateral links' directly spreading activation across the input nodes during single-feature tests and, thereby, equating single-feature performance across dimensions. It appears that for both KRES and people, in the presence of knowledge, learning some features entails learning all of them.

KRES's success and the Perceptron's failure allow us to draw several conclusions about how people acquire concepts. First, it is clear that people do not learn family resemblance categories as collections of isolated sources of information, as the Perceptron assumes. Indeed, KRES's assumption of dimension nonindependence, embodied in its interactive activation properties, accounted for the subjects' highly accurate single-feature classification. This can be interpreted as the single visible feature's reminding people of the missing features, which, in turn, help them select the correct category. The dimensionality effect arises, then, because a greater number of dimensions provides greater opportunity for a given feature to activate other features in memory. Finally, the dimensionality effect increases in the presence of knowledge because features are already strongly associated with one another. Thus, just one visible feature can remind people (now very strongly) of the missing ones. However, knowledge increases the capacity for one feature to evoke the missing features, so that from the categorizers' perspective, one single feature is almost as good as many.

The (non)uniformity of feature learning does not seem to have been a focus of past research on category learning, and it would be interesting to subject other data sets to a similar analysis. Perhaps this finding can help to illumi-

nate the process of category learning and to evaluate other models, whether they include prior knowledge or not.

CONCLUSION

A significant shortcoming of the psychological study of concepts has been that different research programs tend to work on different topics, and the field has done little to integrate these different approaches (Murphy, 2002, chap. 13). In particular, research on structural aspects of category learning and research on the influences of prior knowledge have occupied parallel tracks, and work in one has often had little influence on the other. The present research shows that a formal model of category learning can simultaneously account for a structural phenomenon (the effect of dimensionality), an effect of prior knowledge in category learning, and their interaction. Thus, this work is a step toward devising models that address multiple kinds of phenomena simultaneously. It also shows that although prior knowledge is notoriously difficult to represent, its effects can be modeled computationally by feature-to-feature links.

We reasoned that knowledge and dimensionality influence a common learning mechanism, in which presentation of one feature activates other features related to it. This led us to predict that the two variables would interact, and Experiment 1 documented both their main effects and their interaction in terms of number of dimensions learned. We used KRES to investigate these findings and showed that its interactive activation architecture and representation of knowledge were able to reproduce the observed effects. Prior knowledge magnifies the dimensionality effect in the model, because prior knowledge causes more features to be activated when a single feature is presented. A model lacking interactive activation, the Perceptron, was unable to account for the results even when no knowledge was present. Our analysis revealed that this was largely due to its assumption of independent contribution of a category's properties. KRES, which uses interactive activation during learning and test, was able to model the nonuniform learning of the properties and also the relation between whole-item and single-feature tests.

Although KRES was quite successful, it does not exactly represent the original hypothesis we proposed for the dimensionality effect (Hoffman & Murphy, 2006), which involved more explicit manipulation of attention. There, we suggested that when people are attempting to learn a category, they assume that they should learn most of its properties. Therefore, even when they are classifying items with little error, they allocate free attention to try to learn more properties. In this respect, category learning is different from classical conditioning or other forms of cue learning in which people do not have the belief that most properties should be learned. Indeed, recent work from our lab has shown that an analogue of the classical conditioning blocking effect is not found when people believe they are learning categories but is found when the same problem is construed as predicting the computer's response (Bott, Hoffman, & Murphy, 2007). KRES's success suggests an alternative account of the effect of dimensionality, based

on interactive activation and its learning rule. It is possible that both hypotheses are correct to some degree. However, future research will have to attempt to distinguish them.

AUTHOR NOTE

This work was supported by NIMH Grant MH41704. Correspondence concerning this article should be addressed to A. B. Hoffman, Department of Psychology, University of Texas, 1 University Station A8000, Austin, TX 78712-0187 (e-mail: aaron.hoffman@mail.utexas.edu).

REFERENCES

- ASHBY, G. F., ALFONSO-REESE, L. A., TURKEN, A. U., & WALDRON, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review*, *105*, 442-481.
- BLOOM, P. (2000). *How children learn the meanings of words*. Cambridge, MA: MIT Press.
- BOTT, L., HOFFMAN, A. B., & MURPHY, G. L. (2007). Blocking in category learning. *Journal of Experimental Psychology: General*, *136*, 685-699.
- CAREY, S. (1978). The child as word learner. In M. Halle, J. Bresnan, & G. A. Miller (Eds.), *Linguistic theory and psychological reality* (pp. 264-293). Cambridge, MA: MIT Press.
- COHEN, M. M., & MASSARO, D. W. (1992). On the similarity of categorization models. In F. H. Ashby (Ed.), *Multidimensional models of perception and cognition* (pp. 395-448). Hillsdale, NJ: Erlbaum.
- EDGE, S. E., BRIGHT, R. D., NG, P. C., NOONAN, T. K., & FORD, L. A. (1992). The effect of representation of the processing of probabilistic information. In B. Burns (Ed.), *Percepts, concepts and categories: The representation and processing of information* (pp. 569-601). Amsterdam: Elsevier.
- EDGE, S. E., CASTELLAN, N. J., ROE, R. M., BARNES, J. M., NG, P. C., BRIGHT, R. D., & FORD, L. A. (1996). Irrelevant information in probabilistic categorization. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *22*, 1463-1481.
- ERICKSON, M. A., & KRUSCHKE, J. K. (1998). Rules and exemplars in category learning. *Journal of Experimental Psychology: General*, *127*, 107-140.
- ERICKSON, M. A., & KRUSCHKE, J. K. (2002). Rule-based extrapolation in perceptual categorization. *Psychonomic Bulletin & Review*, *9*, 160-168.
- 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.), *The psychology of learning and motivation* (Vol. 39, pp. 163-199). San Diego: Academic Press.
- HERTZ, J., KROGH, A., & PALMER, R. G. (1991). *Introduction to the theory of neural computation*. Redwood City, CA: Addison-Wesley.
- HOFFMAN, A. B., & MURPHY, G. L. (2006). Category dimensionality and feature knowledge: When more features are learned as easily as fewer. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *32*, 301-315.
- HOPFIELD, J. J. (1987). Learning algorithms and probability distributions in feed-forward and feed-back networks. *Proceedings of the National Academy of Sciences*, *84*, 8429-8433.
- KAMIN, L. J. (1969). Predictability surprise, attention, and conditioning. In B. A. Campbell & R. M. Church (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, & 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.), *The psychology of learning and motivation: Vol. 29. Categorization by humans and machines* (pp. 57-90). San Diego: Academic Press.
- KRUSCHKE, J. K., & JOHANSEN, M. K. (1999). A model of probabilistic category learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *25*, 1083-1119.
- LOVE, B. C., MEDIN, D. L., & GURECKIS, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review*, *111*, 309-332.
- MARKMAN, A. B., & ROSS, B. H. (2003). Category use and category learning. *Psychological Bulletin*, *129*, 592-613.
- MASSARO, D. W., & FRIEDMAN, D. (1990). Models of integration given multiple sources of information. *Psychological Review*, *97*, 225-252.
- MEDIN, D. L., & ORTONY, A. (1989). Psychological essentialism. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 179-195). Cambridge: Cambridge University Press.
- MOVELLAN, J. R., & MCCLELLAND, J. L. (2001). The Morton-Massaro law of information integration: Implications for models of perception. *Psychological Review*, *108*, 113-148.
- MURPHY, G. L. (2001). Fast-mapping children vs. slow-mapping adults: Assumptions about words and concepts in two literatures. *Behavioral & Brain Sciences*, *24*, 1112-1113.
- MURPHY, G. L. (2002). Anti-summary and conclusions. In *The big book of concepts* (pp. 477-498). Cambridge, MA: MIT Press.
- 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.
- NOSOFSKY, R. M., PALMERI, T. J., & MCKINLEY, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review*, *101*, 53-79.
- ODEN, G. C., & MASSARO, D. W. (1978). Integration of featural information in speech perception. *Psychological Review*, *85*, 172-191.
- O'REILLY, R. C. (1996). 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.
- PEARCE, J. M., & BOUTON, M. E. (2001). Theories of associative learning in animals. *Annual Review of Psychology*, *52*, 111-139.
- PITT, M. A., KIM, W., NAVARRO, D., & MYUNG, J. (2006). Global model analysis by parameter space partitioning. *Psychological Review*, *113*, 57-83.
- REDER, L. M., & ROSS, B. H. (1983). Integrating knowledge in different tasks: The role of retrieval strategy on fan effects. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *9*, 55-72.
- REHDER, B., & MURPHY, G. L. (2003). A knowledge-resonance (KRES) model of category learning. *Psychonomic Bulletin & Review*, *10*, 759-784.
- REHDER, B., & ROSS, B. H. (2001). Abstract coherent categories. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *27*, 1261-1275.
- 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* (pp. 64-99). New York: Appleton-Century-Crofts.
- SMITH, E. E., ADAMS, N., & SCHORR, D. (1978). Fact retrieval and the paradox of interference. *Cognitive Psychology*, *10*, 438-464.
- WIDROW, G., & HOFF, M. (1960). Adaptive switching circuits. In *Institute of Radio Engineers, Western Electronic Show and Convention, Convention Record* (Vol. 4, pp. 96-104). New York: Institute of Radio Engineers.

NOTES

1. Pearce and Bouton (2001) proposed that attention is freed after error gets reduced.

2. Equation 1 applies as it is written only when all features predict the same category (e.g., Stimuli 1-5 of condition 5-D₁ in Table 2). For items with an exception feature (e.g., Stimuli 6-10 of condition 5-D₁), one feature predicts the opposite category and, so, effectively "cancels out" another feature. We thus adjusted the formula for such items by subtracting 2 from D_{total} .

3. The failure to fit is not restricted to prototype models. We simulated the ALCOVE model (Kruschke, 1992), a prominent exemplar model of categorization, and found that it also failed to account for the high single-feature response accuracy. In fact, ALCOVE often responded with lower single-feature accuracy than that predicted by the FLMP.

APPENDIX
Filler Items for the Preexposure Task

| | | |
|-----|----------------------|---------------------|
| 1. | Has transparent lens | Has opaque lens |
| 2. | Leather padding | Wool padding |
| 3. | Has tiled floor | Has carpet floor |
| 4. | Holds soda | Holds coffee |
| 5. | Curved edge | Straight edge |
| 6. | Erasable | Indelible |
| 7. | Can be insured | Cannot be insured |
| 8. | Recently built | Built long ago |
| 9. | Water resistant | Water sensitive |
| 10. | Is mechanical | Is organic |
| 11. | High frequency waves | Low frequency waves |
| 12. | Floats in water | Sinks in water |
| 13. | Light activated | Touch activated |
| 14. | Absorbs pollutants | Emits pollutants |
| 15. | Closed with buttons | Closed with snaps |
| 16. | Holds books | Holds magazines |
| 17. | Can take photographs | Can record sound |
| 18. | Hard to operate | Easy to operate |
| 19. | Is comfortable | Is uncomfortable |
| 20. | Above ground | Under ground |

(Manuscript received July 17, 2006;
revision accepted for publication September 6, 2007.)