Psychonomic Bulletin & Review
2003, 10 (4), 759-784

THEORETICAL AND REVIEW ARTICLES

A knowledge-resonance (KRES) model of category learning

BOB REHDER and GREGORY L. MURPHY
New York University, New York, New York

This article introduces a connectionist model of category learning that takes into account the prior knowledge that people bring to new learning situations. In contrast to connectionist learning models that assume a feedforward network and learn by the delta rule or backpropagation, this model, the *knowledge-resonance model*, or *KRES*, employs a recurrent network with bidirectional symmetric connection whose weights are updated according to a contrastive Hebbian learning rule. We demonstrate that when prior knowledge is represented in the network, KRES accounts for a considerable range of empirical results regarding the effects of prior knowledge on category learning, including (1) the accelerated learning that occurs in the presence of knowledge, (2) the better learning in the presence of knowledge of category features that are *not* related to prior knowledge, (3) the reinterpretation of features with ambiguous interpretations in light of error-corrective feedback, and (4) the unlearning of prior knowledge when that knowledge is inappropriate in the context of a particular category.

A traditional assumption in category-learning research, at least since Hull (1920), is that learning is based on observed category members and is relatively independent of other sources of knowledge that the learner already possesses. According to this *data-driven* or *empirical-learning* view of category learning, people associate observed exemplars and the features they display (or a summary representation of those features, such as a prototype or a rule) to the name of the category. Although there now exists a large body of theoretical work that describes how the learning of categories proceeds from the observation of category members, it is also clear that people's knowledge of real-world categories includes more than just the co-occurrence of arbitrary features and category labels. Indeed, recent empirical studies have demonstrated the dramatic influence that a learner's background knowledge often has on the learning process in interpreting and relating a category's features to one another, other concepts, and the category itself (see, Heit, 1997, and Murphy, 1993, 2002, for reviews). The purpose of this article is to present a new computational model of how the acquisition of categories is influenced not only by empirical observations, but also by the prior world knowledge that people bring to the learning task.

Support for this research was provided by funds from the National Science Foundation, Grant SBR 97-20304. We thank Gary Dell, Evan Heit, Lewis Bott, and two anonymous reviewers for their comments on previous versions of this article. Correspondence concerning this article should be addressed to B. Rehder, Department of Psychology, New York University, 6 Washington Place, New York, NY 10003 (e-mail: bob.rehder@nyu.edu).

Murphy (2002) recently concluded that knowledge effects have been found to affect every aspect of conceptual processing for which they have been investigated. For example, prior expectations influence the analysis of a category exemplar into features (Wisniewski & Medin, 1994). Knowledge influences which features are attended to during the learning process and affects the association of features to the category representation (Heit, 1998; Kaplan & Murphy, 2000; Murphy & Allopenna, 1994; Pazzani, 1991; Wisniewski, 1995). In particular, knowledge about causal relations of features can change categorization decisions (Ahn, 1998; Ahn, Kim, Lassaline, & Dennis, 2000; Rehder, 2003a, 2003b; Rehder & Hastie, 2001; Sloman, Love, & Ahn, 1998). People's unsupervised division of items into categories is strongly influenced by their prior knowledge about the items' features (Ahn, 1991; Kaplan & Murphy, 1999; Spalding & Murphy, 1996). Knowledge about specific features can affect the categorization of items after the categories are learned (Wisniewski, 1995), even under speeded conditions with brief stimulus exposures (Lin & Murphy, 1997; Palmeri & Blalock, 2000). Furthermore, structural effects (e.g., those based on feature distribution and overlap) found with meaningless categories may not be found or may even be reversed when the categories are related to prior knowledge (Murphy & Kaplan, 2000; Wattenmaker, Dewey, Murphy, & Medin, 1986). Finally, knowledge effects have been demonstrated to greatly influence category-based induction (Heit & Rubinstein, 1994; Proffitt, Coley, & Medin, 2000; Rehder & Hastie, 2001, in press; Ross & Murphy, 1999).

The amount of evidence for the importance of knowledge in categorization is indeed overwhelming. In fact, its size and diversity suggest that there may not be a single, simple account of how knowledge is involved in conceptual structures and processes. By necessity, the way knowledge is used in the initial acquisition of a category, for example, must be different from the way it is used in induction about a known category. It is an empirical question as to whether the same knowledge structures are involved in different effects, influencing processing in similar ways.

For these reasons, it is critical to explain, at the beginning of a study of knowledge effects, which aspects of knowledge will be examined and (hopefully) explained. The goal of the present study was to understand how knowledge is involved in acquiring new categories through a supervised learning process. Such learning has been the main focus of experimental studies of categories over the past 20 years and has generated the most theoretical development, through such models as prototype theory (Rosch & Mervis, 1975), the context model (Medin & Schaffer, 1978), the generalized context model (Nosofsky, 1986), and various connectionist approaches (e.g., Gluck & Bower, 1988; Kruschke, 1992, 2001; Rumelhart & McClelland, 1986). We will not focus on unsupervised category formation, and other than categorization, we will ignore the use of knowledge in processes that occur after learning has taken place (e.g., the induction of a new property to a category). We will provide only a preliminary analysis of how knowledge might affect logically prior processes, such as the construction of features and the analysis of an item into parts (Goldstone, 2000; Schyns, Goldstone, & Thibaut, 1998; Wisniewski & Medin, 1994). Our hope is that the model we propose can eventually be integrated with accounts of other processes in a way that models that do not include aspects of knowledge would not be. For the present, we focus on the question of how the mental representation of a category results from the combination of empirical knowledge, in the form of observed category exemplars, and prior knowledge about the features of those exemplars. We have tested our account by modeling data from recent studies of knowledge-based concept learning.

We refer to our model of category learning as the *knowledge-resonance model*, or *KRES*. KRES is a connectionist model that specifies prior knowledge in the form of prior concepts and prior relations between concepts, and the learning of a new category takes place in light of that knowledge. A number of connectionist models have been proposed to account for the effects of empirical observations on the formation of new categories, and these models have generally employed standard assumptions, such as feedforward networks (e.g., activation flows only from inputs to outputs) and learning rules based on error signals that traverse the network from outputs to inputs (e.g., the delta rule, backpropagation; Gluck & Bower, 1988; Kruschke, 1992, 2001). To date, attempts to incorporate the effects of prior knowledge into connectionist models have been restricted to extensions of this same basic architecture (e.g., Choi, McDaniel, & Busemeyer, 1993; Heit & Bott, 2000). KRES departs from these previous attempts in its assumptions regarding both activation dynamics and the propagation of error. First, in contrast to feedforward networks, KRES employs *recurrent networks* in which connections among units are bidirectional and activation is allowed to flow not only from inputs to outputs, but also from outputs to inputs and back again. Recurrent networks respond to input signals by each unit's iteratively adjusting its activation in light of all the other units until the network *settles*—that is, until change in the units' activation levels ceases. This settling process can be understood as an interpretation of the input in light of the knowledge or constraints that are encoded in the network. As applied to the categorization problems considered here, a KRES network accepts input signals that represent an object's features and interprets (i.e., classifies) that object by settling into a state in which the object's category label is active.

Second, rather than backpropagation, KRES employs *contrastive Hebbian learning* (CHL) as a learning rule applied to deterministic networks (Movellan, 1989). Backpropagation has been criticized as being neurally implausible, because it requires nonlocal information regarding the error generated from corrective feedback in order for connection weights to be updated (Zipser, 1986). In contrast, CHL propagates error, using the same connections as those that propagate activation. During an initial *minus phase*, a network is allowed to settle in light of a certain input pattern. In the ensuing *plus phase*, the network is provided with error-corrective feedback by being presented with the output pattern that should have been computed during the minus phase and is allowed to resettle in light of that correct pattern. After the plus phase, connection weights are updated as a function of the difference between the activation of units between the two phases. O'Reilly (1996) has shown that CHL is closely related to the pattern-learning *recirculation algorithm* proposed by Hinton and McClelland (1988). Its performance is also closely related to a version of backpropagation that accommodates recurrent connections among units (Almeida, 1987; Pineda, 1987), despite the absence of a separate network that propagates error.

In addition to activation dynamics and learning, the third central component of KRES is its representation of prior knowledge. As for any cognitive model that purports to represent real-world knowledge, we were faced with the problem that knowledge representation is still one of the less understood aspects of cognitive psychology. For example, although progress has been made in developing representations necessary to account for the structured nature of some kinds of world knowledge (e.g., schemata and taxonomic hierarchies), there is little agreement on the overall form of representation of complex domains, such as biology, American politics, personalities, and so on. Nevertheless, even without a complete theory of knowledge representation, we believe that a useful model of knowledge effects can be de-

veloped, as long as the essential influences of prior knowledge on category learning is somehow captured.

With this goal in mind, our method of representing prior knowledge in KRES includes two somewhat different approaches. The idea behind the first approach is to relate or constrain pairs of features by linking them with feature-to-feature connections. The assumption is that features that are related through prior knowledge will have preexisting excitatory connections relating them, that features that are inconsistent will have inhibitory connections, and that features that are not involved in any common knowledge structures will have no such links (or links with zero weight). Our claim is that, at least for the purposes of modeling the learning of new categories, feature-to-feature connections can approximate the effect of a number of different types of pairwise semantic relations, including causal relations, function–form relationships, part–whole relationships, feature co-occurrence, and so on.

The second approach to representing knowledge is borrowed from Heit and Bott (2000). The notion here is that some category learning is based in part on the similarity of the new category to a known category. For example, when consumers learned about DVD players, they no doubt used their knowledge of videocassette recorders, which served a similar function, and CD players, which used a similar technology, in order to understand and learn about the new kind of machine. Heit and Bott accounted for such knowledge by including in the network prior concepts that had some of the same features as the to-be-learned categories. Although we agree that this is one source of knowledge, we also believe that it is somewhat limited in what it can accomplish. For example, a number of experiments on knowledge effects (described below) have used features that are related to one another but that do not correspond to any existing category. Thus, we incorporate prior concepts as one source of knowledge but add feature–feature connections in order to more flexibly represent knowledge.

Our use of these two relatively simple forms of knowledge should not be interpreted as ruling out the existence and importance of other, more complex forms. For example, as has already been mentioned, KRES does not explicitly represent schemata or taxonomic hierarchies (e.g., Brachman, 1979; Brewer & Nakamura, 1984; Rumelhart, 1980). Also, it does not represent propositional knowledge of the form that requires binding concepts to their roles as arguments of predicates (e.g., Fodor & Pylyshyn, 1988; Hummel & Holyoak, 1997; Marcus, 2001). It also does not represent specific prior examples or cases from preexisting categories that might be accessed by similarity or analogy (as proposed, for example, by Heit's, 1994, *integration model*). In the General Discussion section, we will assess the importance of these other forms of knowledge on category learning and will consider ways of incorporating some of them into later versions of the model. We will pay special attention to comparing KRES's assumptions regarding the representation of knowledge with those of the integration model, which has simulated some of the same empirical studies as those we will present here.

We will now describe the KRES model in detail, including a description of its activation dynamics, learning algorithm, and representation of knowledge. We then will report the results of several simulations of empirical category-learning data. We will demonstrate that KRES is able to account for a number of striking empirical category learning results when prior knowledge is present, including (1) the accelerated learning that occurs in the presence of knowledge, (2) the learning of category features that are *not* related to the prior knowledge, (3) the reinterpretation of ambiguous features in light of corrective feedback, and (4) the unlearning of prior knowledge when that knowledge is inappropriate in the context of a particular category. These results will be attributed to three distinguishing characteristics of KRES: (1) a recurrent network that allows category features to be interpreted in light of prior knowledge, (2) a recurrent network that allows activation to flow from outputs to inputs, and (3) the CHL algorithm that allows (re)learning of all connections in a network, including those that represent prior knowledge.

THE KNOWLEDGE-RESONANCE MODEL (KRES)

Two examples of a KRES model are presented in Figures 1 and 2. In these figures, circles depict *units* that represent category labels (X and Y), category features (A_0, A_1, B_0, B_1, etc.), or prior concepts (P_0 and P_1). To simplify the depiction of connections among groups of units, units are organized into *layers*, specified by boxes. Units may belong to more than one layer, and layers may intersect and contain other layers. Solid lines among layers represent connections among units provided by prior knowledge. Solid lines terminated with black circles are excitatory connections; those terminated with hollow circles are inhibitory connections. Dashed lines represent new, to-be-learned connections. By default, two connected layers are fully connected (i.e., every unit is connected to every other unit), unless annotated with "1:1" (i.e., one to one), in which case each unit in a layer is connected to only one unit in the other layer. Finally, double dashed lines represent external perceptual inputs. As will be described below, both the feature units and the category label units receive external input, although at different phases of the learning process.

Representational Assumptions

A unit has a level of activation in the range 0 to 1 that represents the activation of the concept. A unit i's activation, act_i, is a sigmoid function of its total input, that is,

$$act_i = 1 / [1 + \exp(-total\text{-}input_i)], \quad (1)$$

and its total input comes from three sources:

$$total\text{-}input_i = net\text{-}input_i + external\text{-}input_i + bias_i. \quad (2)$$

Network input represents the input received from other units in the network. External input represents the presence of (evidence for) the feature in the external environment. Finally, each unit has its own bias, which determines how easy or difficult it is to activate the unit. A unit's bias can be interpreted as a measure of the prior probability that the feature is present in the environment. Each of these inputs is a real-valued number.

Relations between concepts are represented as connections with a real-valued weight, $weight_{ij}$, in the range minus to plus infinity. Connections are constrained to be symmetric—that is, $weight_{ij} = weight_{ji}$.

A unit's network input is computed by multiplying the activation of each unit to which it is connected by the connection's weight and then summing over those units:

$$net\text{-}input_i = \bullet_j \, act_j * weight_{ij}. \quad (3)$$

In many applications, two (or more) features might be treated as mutually exclusive values on a single dimension, often called *substitutive features.* In Figure 1, the stimulus space is assumed to consist of five binary-valued dimensions, with A_0 and A_1 representing the two values on dimension A, B_0, and B_1 representing the two values on dimension B, and so on. To represent the mutual exclusivity constraint, there are inhibitory connections between units that represent the 0 value on a dimension and the units that represent the corresponding 1 value. In Figures 1 and 2, the units that represent prior concepts (P_0 and P_1) and the to-be-learned category labels (X and Y) are also assumed to be mutually exclusive and, hence, are linked by an inhibitory connection. Note that KRES departs from many connectionist models of concepts (e.g., Anderson & Murphy, 1986; Estes, 1994; Heit & Bott, 2000; Kruschke, 1992; McClelland & Rumelhart, 1985) by representing binary dimensions with two units, rather than with a single unit that takes on the value −1 or +1. This approach allows mutually exclusive features to be involved in their own network of semantic relations. For example, unlike the traditional approach, KRES can represent that white and red are mutually exclusive, that white, but not red, is related to purity, and that red, but not white, is related to communism.

The Representation of Prior Knowledge

As has been described earlier, KRES represents prior knowledge in the form of known concepts (i.e., units) and/or prior associations (i.e., connections) between units. In Figure 1, P_0 is a prior concept related to features A_0, B_0,

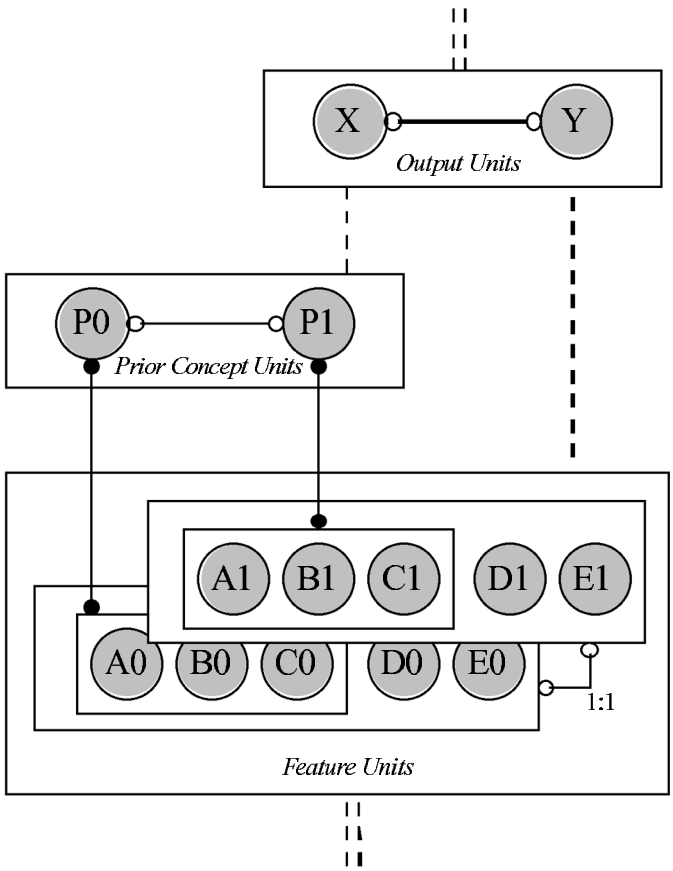

Figure 1. A KRES model with prior concept units.

and C_0, and P_1 is a prior concept related to features A_1, B_1, and C_1. The relations between features and prior concepts are rendered as excitatory connections between the units.

Prior knowledge may also be represented in the form of direct excitatory connections among the features, as is shown in Figure 2. In Figure 2, it is assumed that features A_0, B_0, and C_0 are related by prior knowledge, as are features A_1, B_1, and C_1. These relations link the features directly (e.g., wings are associated with flying), rather than through a prior concept.

In the simulations that will follow, we have employed either prior concept units or direct interfeature connections in modeling the prior knowledge of category learners. Although the choice of which of these two forms of representations to use can be somewhat arbitrary in certain cases (i.e., based on our own intuitions regarding the form of the prior knowledge involved), it should be noted that both have a similar overall effect on learning: As the result of these mutually excitatory connections in a recurrent network, units achieve a higher activation level than they would otherwise, and this greater activation leads to faster learning, as will be described below.

Classification via Constraint Satisfaction

Before KRES is presented with external input that represents an object's features, the activation of each unit is initialized to a value determined solely by its bias (i.e., the activation of each unit is initialized to the prior probability that it is present). The external input of a feature unit is then set to 1.0 if the feature is present in the input, −1.0 if it is absent, and 0.0 if its presence or absence is unknown. The external input of all other units is set to 0.0. The model then undergoes a standard multicycle constraint satisfaction process that involves updating the activation of each unit in each cycle in light of its external input, its bias, and its current network input. (In each cycle, the serial order in which units are updated is determined by randomly sampling units without replacement.)[1] After each cycle, the *harmony* of the network is computed (Hinton & Sejnowski, 1986; Hopfield, 1982; Smolensky, 1986):

$$harmony = \bullet_i \bullet_j act_i * act_j * weight_{ij}. \quad (4)$$

Constraint satisfaction continues until the network settles, as indicated by a change in harmony from one cycle to the next of less than 0.00001.

In this article we present simulations of the results of several empirical studies, in which KRES was used to model two dependent measures: response times (RTs) and error rates. The number of cycles required for the network to settle was assumed to correspond to RT. Error rates were modeled by assuming that the activation values associated with the category label units X and Y that obtain after the network settles represent the evidence that the current input pattern should be classified as an X and a Y, respectively. These activation values were mapped into a categorization decision in the standard way, following Luce's choice rule:

$$choice\text{-}probability\ (X, Y) = act_X / (act_X + act_Y). \quad (5)$$

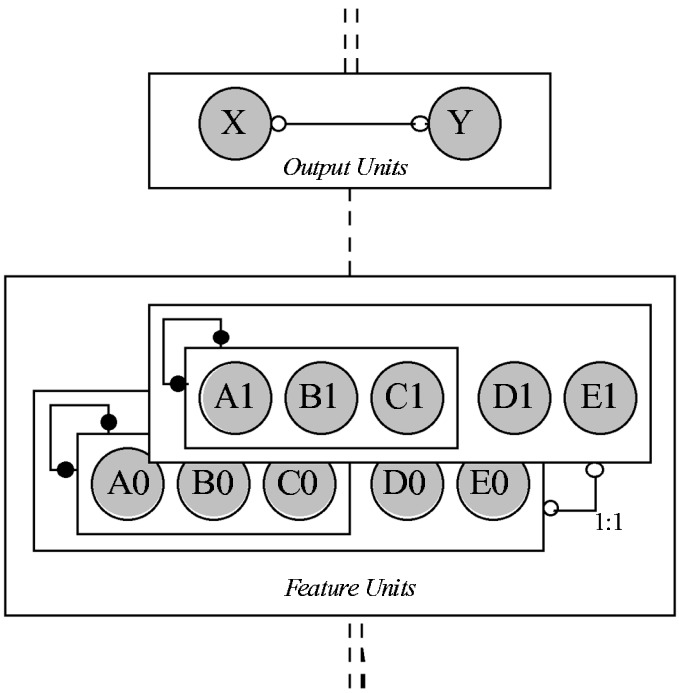

Figure 2. A KRES model with interfeature connections.

Contrastive Hebbian Learning (CHL)

As has been described earlier, the settling of a network that results as a consequence of presenting just the feature units with external inputs is referred to as the minus phase. In the plus phase, error-correcting feedback is provided to the network by setting the external inputs of the correct and incorrect category label units to 1.0 and −1.0, respectively, and allowing the network to resettle in light of these additional external inputs. We refer to the activation values of unit i that obtain after the minus and plus phases as act_i^- and act_i^+, respectively. After the plus phase, the connection weights are updated according to the CHL rule:

$$\varnothing weight_{ij} = lrate * (act_i^+ * act_j^+ - act_i^- * act_j^-), \quad (6)$$

where *lrate* is a learning rate parameter. Because $act_i^- * act_j^-$ and $act_i^+ * act_j^+$ are the derivatives with respect to $weight_{ij}$ of the harmony function (Equation 4) in the minus and plus phases, respectively, this learning rule can be interpreted as having the effect of increasing network harmony in the plus phase and decreasing it in the minus phase, making it more likely that the network will settle into a state of activation more closely associated with the plus phase when the training pattern is re-presented in a subsequent training trial (Movellan, 1989). O'Reilly (1996) has shown that CHL is related to the Almeida–Pineda version of backpropagation for recurrent networks but that CHL achieves faster learning because it constrains weights to be symmetric and incorporates a simple numerical integration technique that approximates the gradient of the error derivative. We will demonstrate in Simulation 1 how CHL approximates the delta rule for a simple one-layer network at the early stages of learning when the effect of recurrent connections is minimal.

Network Training

Before a KRES network is trained, all connection weights are set to their initial values. All new, to-be-learned connections are initialized to a random value in the range [−0.1, 0.1], and the biases of all the units are initialized to 0. The weights of those excitatory and inhibitory connections that represent prior knowledge are initialized to a value that differs across simulations (as specified below) and do not change during category learning.

As in the behavioral experiments we simulated, training consists of repeatedly presenting a set of training examples in blocks, with the order of the training patterns randomized within each block. Training continues either for a fixed number of blocks or until the average error for a training block falls below an error criterion. The average error associated with a block is computed by summing the errors associated with each training pattern in the block and dividing by the number of training patterns. The error associated with a training pattern is calculated by computing the squared difference between the activation levels of the category label units and their correct values (0 or 1) and summing these squared differences over the two category label units.

KRES SIMULATION OF EMPIRICAL DATA

The following sections present KRES simulations of six empirical data sets. The learning rate and error criterion varied across simulations. In each simulation, the KRES model was run 100 times with a different random set of initial weights, and the results reported below were averaged over those 100 runs.

Simulation 1: Prototype Effects and Cue Competition

The primary purpose of KRES is to account for the effect of prior knowledge on category learning. In this initial simulation, however, we show that KRES exhibits some properties that make it a candidate model of category learning in the absence of knowledge. In particular, we have shown that KRES exhibits both prototype effects and cue competition effects, such as overshadowing and blocking.

Since the popularization of the notion of probabilistic categories in the 1970s, it has usually been found that category membership is related directly to the number of typical features that an object displays, where typical features are those that appear frequently among category members and seldom among members of other categories (Hampton, 1979; Rosch & Mervis, 1975; Smith & Medin, 1981). For example, Rosch and Mervis constructed family resemblance categories based on alphanumeric characters. Some characters occurred frequently in the category, and some occurred less frequently. Also, some characters occurred more frequently in contrast categories, and others occurred less frequently. Rosch and Mervis demonstrated that items were classified more accurately if they possessed features common to the category but not features that occurred in contrast categories. Many other studies have shown experimentally that the category prototype is classified accurately, even if it has not been seen before (e.g., Franks & Bransford, 1971; Posner & Keele, 1968).

This sort of demonstration is very important, because typicality effects are by far the most frequent empirical phenomenon found in studies of concepts (Murphy, 2002) and the clearest demonstrations of typicality have been in studies in which no knowledge was involved (e.g., Rosch & Mervis's, 1975, alphanumeric characters and Posner & Keele's, 1968, dot patterns). Furthermore, typicality effects in natural categories can be largely, although not entirely, explained by structural factors (Barsalou, 1985). Therefore, we wished to demonstrate that the basic KRES architecture would exhibit the usual typicality gradient based on purely structural factors, before going on to explore knowledge effects.

To determine whether KRES would exhibit typicality effects, we trained it on the exemplars presented in Table 1. The exemplars consist of five binary-valued substitutive features, where 1 and 0 represent the two values on a single dimension. Note that although dimension value 1 is typical of Category *X* and 0 is typical of Category *Y*, no exemplar contains all the features typical of one category.

Table 1
Training Exemplars for Simulation 1

Dimension					Category
A	B	C	D	E	Label
1	1	1	1	0	*X*
1	1	1	0	1	*X*
1	1	0	1	1	*X*
1	0	1	1	1	*X*
0	1	1	1	1	*X*
0	0	0	0	1	*Y*
0	0	0	1	0	*Y*
0	0	1	0	0	*Y*
0	1	0	0	0	*Y*
1	0	0	0	0	*Y*

That is, during training, the prototypes of Categories *X* and *Y* were never presented. This sort of factorial structure has been used in many category-learning studies, since it ensures that no feature is either necessary or sufficient for categorization.

This KRES model was like those shown in Figures 1 and 2, with an inhibitory connection of −2.0 between features on the same dimension, but without either prior concepts or interfeature connections, since the features were assumed to be arbitrary. Training proceeded with a learning rate of 0.10 until an error criterion of 0.10 was reached. After training, the model was tested with all possible combinations of the five binary dimensions. Figure 3 presents KRES's choice probabilities as a function of the number of features typical of Category *X* present in the test pattern. As Figure 3 demonstrates, the Category *X* prototype 11111 was classified more accurately as an *X* than were the original *X* training exemplars (i.e., those that possessed four out of five typical *X* features; see Table 1), even though it was never seen. Likewise, the Category *Y* prototype 00000 was classified more accurately as a *Y* than were the original *Y* training exemplars. That is, KRES exhibited classic typicality effects. The borderline items, containing only three features of a single category (out of five) were generally classified correctly, but less often than the more typical ones.

With a simple modification, the set of training exemplars shown in Table 1 can also be used to demonstrate one of the cue competition effects known as *overshadowing* (Gluck & Bower, 1988; Kamin, 1969). According to standard accounts of associative learning, cues compete with one another so that the presence of stronger cues will result in weaker cues being less strongly associated to the outcome. To simulate this effect, an additional dimension, *F*, was added to the training exemplars presented in Table 1 that was perfectly predictive of category membership: Whenever an exemplar had a 1 on dimension *F*, it belonged to Category *X*; whenever it had a 0, it belonged to *Y*.

A KRES model with the same parameters was run on this new training set. As was expected given the presence of the perfectly predictive Dimension *F*, the error criterion was reached in fewer blocks in this second simulation (8.0) than in the original one (10.1). Moreover, the results indicated that the features on Dimensions *A–E* were not learned as well. First, the connection weights between those features and their correct category label were reduced from an average of 0.634 without the presence of Dimension *F* to an average of 0.461 with it. Second, as a result of these weaker associations, the activation of the correct category label unit was reduced when the network was tested with single features. To test the network with a single feature, the unit representing that feature was given an external input of 1, the unit representing the other feature on the same dimension was given an input of −1, and all other units were given 0. Whereas the choice probability associated with individual features on Dimensions *A–E* was .81 in the original simulation, it was reduced to .73 in the presence of Dimension *F*. That is, Dimension *F* overshadowed the learning of the other features. Because of the error-driven nature of the CHL rule, it is straightforward to show that KRES networks also exhibit standard *blocking effects*, in which feature-to-category associations that are already learned prevent the learning of new associations.

These initial simulations demonstrate that despite its nonstandard activation dynamics (recurrent networks) and learning rule (CHL), KRES can learn categories and exhibits standard prototype and cue competition effects. The fact that KRES exhibits these effects is not surprising, because it can be shown that for the simple network employed in Simulation 1, the CHL rule approximates the delta rule. Two assumptions are necessary to show this. First, assume that during the plus phase of the CHL procedure, the correct and incorrect category labels take on the values that they should ideally reach in the presence of the input pattern (namely, 1 and 0), rather then just having their external inputs set to 1 and −1, respectively.[2] Second, during the early parts of learning, connection weights are close to zero. As a result, during the plus phase, the new activation values of the category

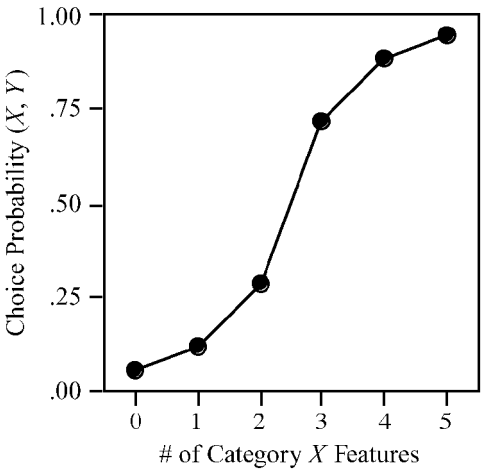

Figure 3. Classification test results from Simulation 1.

label units return little activation to the feature units, and hence, the activation values of the feature units change only little between the plus and the minus phases. In other words, early in learning, $act_i^+ \cong act_i^- = act_i$ for feature unit i. Under these conditions, the CHL rule Equation 6 becomes

$$\begin{aligned} \varnothing weight_{ij} &= lrate * (act_i * act_j^+ - act_i * act_j^-) \\ &= lrate * act_i * (act_j^+ - act_j^-), \end{aligned} \quad (7)$$

where i is an input (feature) unit and j is an output (category label) unit. Because act_j^+ is the "target" activation value for the output unit (0 or 1), Equation 7 is simply the delta rule.

Our central purpose in this article is to show that KRES is able to account for a variety of knowledge-related learning effects that have, until now, stood beyond the reach of traditional empirical models of category learning. As will be seen (most clearly in Simulations 4 and 5), one of the mechanisms by which this is accomplished is by the adjustment of the activation of the feature units. For example, when features are involved in networks of excitatory connections that represent prior knowledge, the result is that those features attain higher activation levels, as represented by act_i in Equation 7. As act_i increases, Equation 7 indicates that the rate at which features are associated to category label units increases (i.e., learning is faster).

At the same time, an equally important goal is to show that by being grounded in a learning algorithm with close connections to the delta rule (and for multilayer networks, backpropagation), KRES is also a member of the family of empirical-learning models that have been shown to exhibit a number of phenomena of human associative learning, such as prototype effects and cue competition. The result is a model that uses prior knowledge during learning while simultaneously carrying out associative learning. As will be seen, this feature of KRES is crucial for accounting for the human learning data.[3]

Simulation 2: Learning With Prior Concepts

In the literature on category learning with prior knowledge, perhaps the most pervasive effect is that learning is dramatically accelerated when prior knowledge is consistent with the empirical structure of training exemplars. For example, Wattenmaker et al. (1986, Experiment 1, linearly separable condition) presented examples of two categories whose features either could be (related condition) or could not be (unrelated condition) related to an underlying theme or trait. (The related and the unrelated conditions were referred to as the *trait* and the *control* conditions by Wattenmaker et al.[4]) For instance, in the related condition, one category had four typical features that could be related to the trait *honesty* (e.g., "returned the wallet he had found in the park," "admitted to his neighbor that he had broken his rake," "told the host that he was late for the dinner party because he had overslept," etc.), whereas the other category had four typical features that could be related to the trait *dishonesty* or *tactfulness* (e.g., "pretended that he wasn't bothered when a kid threw a Frisbee and knocked the newspaper out of his hands," "told his visiting aunt that he liked her dress even though he thought it was tasteless," etc.). In the unrelated condition, the four typical features of each category could not be related to any common theme. During training, Wattenmaker et al. presented learners with category examples that contained most, but not all, of the features typical of the category (as in our Simulation 1). They found that subjects reached a learning criterion in many fewer blocks in the related condition (8.8) than in the unrelated condition (13.7), a result they attributed to learners' relating the features to the trait in the former condition, but not in the latter.

This experiment was simulated by a KRES model like the one shown in Figure 1, with eight features representing the two values on four binary dimensions. In the related, but not in the unrelated, condition, the four features with the 0 dimension value had excitatory connections to a prior concept unit, and the four features with the 1 dimension value had excitatory connections to a different prior concept unit. The weight on these excitatory connections was set to 0.75, the weight on inhibitory connections was set to −2.0, the learning rate was 0.15, and the error criterion was 0.10. We used prior concept units in this simulation, because it seemed clear that the subjects already had concepts corresponding to the two traits Wattenmaker et al. (1986) used (i.e., *honesty* and *dishonesty*).

Figure 4 presents the results from Wattenmaker et al. (1986), along with the KRES simulation results (aver-

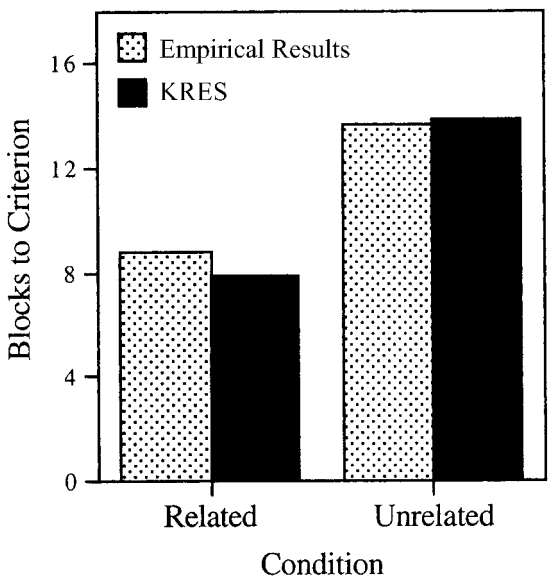

Figure 4. Results from Wattenmaker, Dewey, Murphy, and Medin (1986), Experiment 1, linearly separable condition, and from Simulation 2.

aged over 100 runs, as was explained earlier). As this figure shows, KRES replicated the basic learning advantage shown when a category's typical features could be related to an underlying trait or theme. That is, the KRES model reached its learning criterion in many fewer blocks when the categories' features were connected to a prior concept than when they were not.

KRES produced a learning advantage in the related condition because, on each training trial, the training pattern tended to activate its corresponding prior concept unit. Figure 5A shows the average activation of the features of the correct category during each training trial for both the related and the unrelated conditions, as well as the activation of the prior concept units that were activated by the training pattern in the related condition. The figure indicates that in the related condition, the feature units activated the prior concepts units to which they were connected. Because the correct prior concept units were activated on every training trial, the connection weights between the prior concepts and the category label units grew quickly, as is shown in Figure 5B. In comparison, the connection weights between the features and the category labels grew more slowly. This occurred because each feature appeared with an exemplar from the wrong category on some trials of each training block, decrementing the connection weight between the feature and its correct category node. It is the constant conjunction of the prior concepts and the category labels that was mostly responsible for faster learning in the related condition.

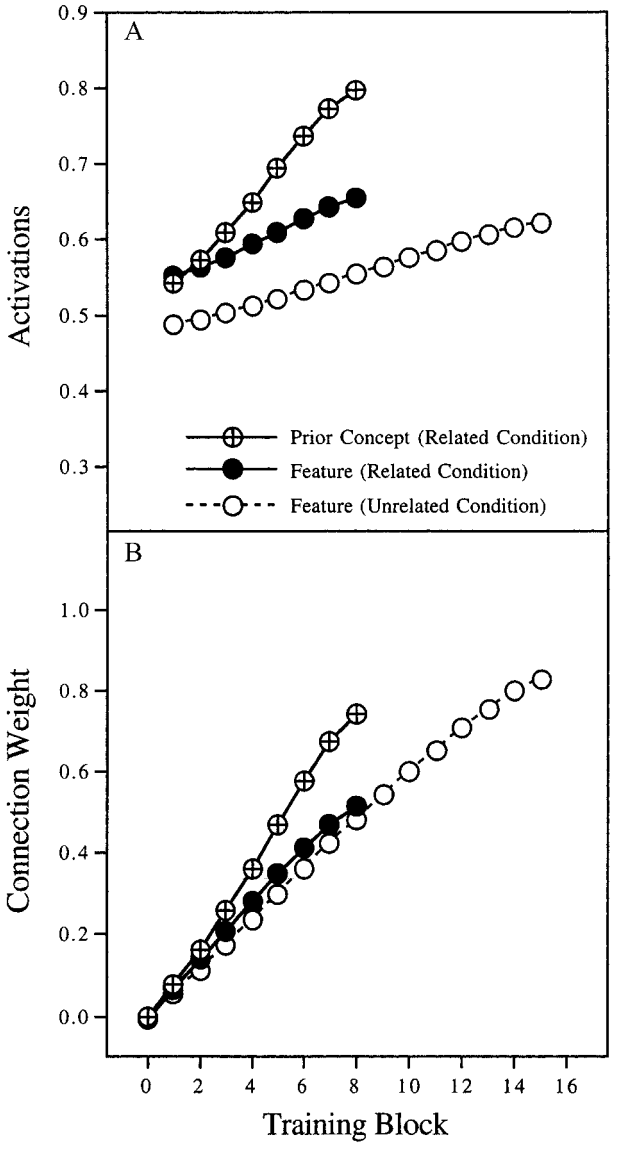

Figure 5. Results from Simulation 2. (A) Average activation values. (B) Average weights to the correct category label units.

Three other aspects of Figure 5 demonstrate properties of KRES's activation dynamics. First, the activation of feature units was greater in the related than in the unrelated condition. This occurred because the feature units receive recurrent input from the prior concept unit that they activated. The result was somewhat faster learning of the weights on the direct connections between the features and the category labels in the related than in the unrelated condition (Figure 5B). Second, the activation levels of the feature units in the related and the unrelated conditions and of the prior concept units in the related condition tended to become larger as training proceeded. This occurred because once positive connections to the category labels were formed, the category labels recurrently sent activation back to these units. This effect was strongest for the prior concept units, which had the strongest connections to the category labels. This further accelerated learning in the related condition in the later stages of learning. Finally, at the end of training, the connection weights to category labels were larger in the unrelated condition than in the related condition. This result might seem puzzling, because the same error criterion was used in both conditions and one might expect the same connection weights at the same level of performance. This difference in connection weights occurred because, whereas the category label units were activated by both feature and prior concept units in the related condition, they were activated by only feature units in the unrelated condition. The result was that the unrelated condition required greater connection weights from the input to attain the same activation of the category labels as that achieved in the related condition. This difference is analogous to the cue competition effect shown in Simulation 1: Because the prior concept units aided performance, the connection weights between input features and category labels were not as large.

Simulation 3: Learning Facilitated by Knowledge

Simulation 2 provided a basic demonstration of the fact that knowledge speeds category learning when category features can be related to a common theme. Heit

and Bott (2000) conducted a more detailed study of category learning in the presence of a prior theme by employing categories where some, but not all, of the features could be related to the theme. Heit and Bott created two categories with 16 features each, 8 of which could be related to an underlying theme and 8 of which could not. For example, for the category whose underlying theme was *church building*, some of the related features were "lit by candles," "has steeply angled roof," "quiet building," and "ornately decorated." Some of the unrelated features were "near a bus station" and "has gas central heating." The subjects were required to discriminate examples of church buildings from examples of office buildings (although, of course, the categories were not given these labels), with related features such as "lit by fluorescent lights" and "has metal furniture" and unrelated features such as "not near a bus station" and "has electric central heating." (Each exemplar also possessed a small number of idiosyncratic features, which we will not consider.)

In order to assess the time course of learning, Heit and Bott (2000) presented test blocks after each block of training, in which subjects were required to classify related and unrelated features presented alone. Because these investigators were also interested in how subjects would classify previously unobserved features, a small number of the related and unrelated features were never presented during training.

The subjects were trained on a fixed number of training blocks. The results, averaged over Heit and Bott's (2000) Experiments 1 (church vs. office buildings) and 2 (tractors vs. racing cars), are presented in Figure 6. The figure shows the percentage of correct classifications of individual features in the test blocks as a function of the number of blocks of training and the type of features. Several things should be noted. First, the subjects learned the presented related features better than the presented unrelated features. Second, they correctly classified those related features that were *never presented* in training examples. Third, despite the presence of the theme, the subjects still exhibited considerable learning of those unrelated features that were presented. Finally, as was expected, the subjects were at chance on those unrelated features that were not presented.

This experiment was simulated by a KRES model with 32 features representing the two values on 16 binary dimensions. Eight features with the 0 dimension value (e.g., "lit by candles") were provided excitatory connections to a prior concept unit (the *church building* concept), and the corresponding eight features with the 1 values on the same dimensions (e.g., "lit by fluorescent lights") were provided excitatory connections to the other prior concept (the *office building* concept). The remaining 16 features (2 on 8 dimensions) had no links to the prior concepts. The weight on the excitatory connections among features was set to 0.65, the weight on inhibitory connections was set to −2.0, the learning rate

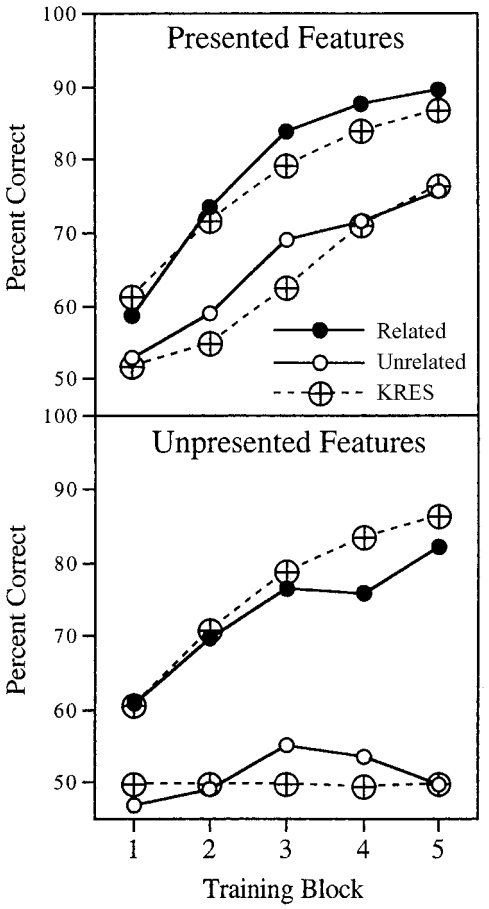

Figure 6. Results from Heit and Bott (2000), Experiments 1 and 2, and from Simulation 3.

was 0.15, and the error criterion was 0.10. Like the subjects in Heit and Bott (2000), the model was run for a fixed number of training blocks (five). After training, the model was tested by being presented with single features, as in Simulation 1.

The results of KRES's single-feature tests are presented in Figure 6, superimposed on the empirical data. The figure shows that KRES reproduced the qualitative results from Heit and Bott (2000). First, KRES classified presented related features more accurately than presented unrelated features. This occurred for the same reasons as in Simulation 2. During learning, the prior concept units were activated on every training trial and, hence, quickly became strongly associated to one of the category labels. During test, the presented related, but not the unrelated features, activated their correct prior concept unit, which then activates the correct label. As a result, the related features were classified more accurately than the unrelated ones.

Second, KRES classified unpresented related features accurately, because these features also activated the prior concept unit to which they were (preexperimentally) re-

lated, which in turn activated the unit for the correct category. For example, before the experiment, Heit and Bott's (2000) subjects already knew that churches are often built out of stone. After the training phase of the experiment, they also knew that one of the experimental categories was related to church buildings (e.g., "Category A is a house of worship of some kind"). Therefore, when asked which experimental category the feature "built of stone" was related to, they picked Category A, because (according to KRES) the built-of-stone feature node activated the church concept, which then activated Category A. This accurate categorization occurred even though none of the examples of Category A presented during the experiment was described as being built out of stone.

Third, KRES exhibited considerable learning of the presented unrelated features. In Simulation 1, we saw that KRES could perform associative learning of the sort necessary to acquire new concepts that did not involve prior knowledge. In this simulation, we see that KRES could *simultaneously* perform empirical learning of features unrelated to prior knowledge and the more knowledge-based learning of related features. That is, learners did not focus solely on the prior concepts ("Category A is a house of worship of some kind") but also learned properties that were not related by prior knowledge to the concepts ("Members of Category A are usually near bus stations"). The model learned both.

Finally, KRES exhibited no learning of the unpresented unrelated features, revealing that the model does not have ESP.

Simulation 4: Prior Knowledge Without Prior Concepts

Although the empirical results reported in the previous two sections provide evidence for the importance of prior knowledge during category learning, it is arguable whether the learning that took place actually consisted of learning new categories. The subjects already knew such concepts as *honesty* (in Simulation 2) and *church building* (in Simulation 3), and it might be argued that most of the learning that took place was merely to associate these preexisting categories with new category labels (although perhaps refined with some additional features). Indeed, the KRES simulations of these data explicitly postulated the presence of units that represented these preexisting concepts.

Because of the use of prior concept units, it can also be shown that the success of Simulations 2 and 3 did not depend critically on the distinctive features of KRES, such as recurrent networks and contrastive Hebbian learning. Heit and Bott (2000) have proposed a feedforward connectionist model called *Baywatch*, which learns according to the delta rule. As we assumed in Simulations 2 and 3, Heit and Bott suggested that features activate prior concepts, which are then directly associated to the new category labels. Unlike with KRES, however, in Baywatch those prior concepts do not return activation to the feature units. Heit and Bott demonstrated that Baywatch reproduces the pattern of empirical results shown in Figure 6, despite the absence of such recurrent connections.

As was discussed earlier, there is no doubt that the learning of some new categories benefits from their similarity to familiar categories. In such cases, prior concept nodes, or something like them, may well be involved and may aid learning. However, in other cases, a new category may be generally consistent with knowledge but may not correspond precisely, or even approximately, to any particular known concept. That is, some new concepts may "make sense" in terms of being plausible or consistent with world knowledge and, therefore, may be easier to learn than those that are implausible, even if they are not themselves familiar. For such cases, a different approach seems to be called for.

The empirical study of Murphy and Allopenna (1994, Experiment 2) may be such a case. Subjects in a related condition were asked to discriminate two categories that had six features that could be described as coming from two different themes: *arctic vehicles* ("drives on glaciers," "made in Norway," "heavily insulated," etc.) or *jungle vehicles* ("drives in jungles," "made in Africa," "lightly insulated," etc.). Each category exemplar also possessed features drawn from three dimensions that were unrelated to the other features (e.g., "four door" vs. "two door," "license plate on front" vs. "license plate on back") and that were not predictive of category membership. The learning performance of these subjects was compared with those in an unrelated control condition in which the same features were recombined in such a way that they no longer described a coherent category. (The related and the unrelated conditions were referred to as the *theme* and the *no-theme* conditions by Murphy and Allopenna.) As in Wattenmaker et al.'s (1986) study presented above, related subjects reached a learning criterion in fewer blocks (2.5) than did those in the unrelated control condition (4.1). Unlike in Wattenmaker et al. and Heit and Bott (2000), however, the categories employed by Murphy and Allopenna were rated as novel, as compared with the control categories, by an independent group of subjects (see also Spalding & Murphy, 1999). Thus, the prior concept nodes used in Simulation 2 would not be appropriate here.

To simulate these results without assuming prior knowledge of the concepts *arctic vehicle* and *jungle vehicle*, we created a KRES model like the one shown in Figure 2 that assumed the presence of prior knowledge only in the form of connections between features—that is, there were no prior concept nodes. The model used 18 features, representing the two values on nine binary dimensions. In the related, but not in the unrelated, condition, 6 features with the 0 dimension value were interrelated with excitatory connections, as were the corresponding 6 features with the 1 dimension value. The weight on these excitatory connections was initialized to 0.55, the weight on inhibitory connections was set to −2.0, the learning rate was set to 0.125, and the error criterion was set to 0.05.

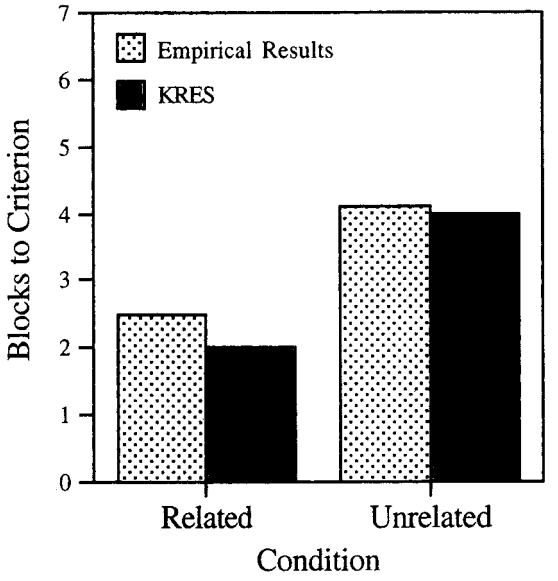

Figure 7. Results from Murphy and Allopenna (1994), Experiment 2, and from Simulation 4.

The number of blocks required to reach criterion as a function of condition are presented in Figure 7 for both the experimental subjects and KRES. As the figure indicates, KRES reproduced the learning advantage found in the related condition. Since there were no prior concept nodes in this version of the model, this advantage can be directly attributed to KRES's use of recurrent networks: The mutual excitation of knowledge-related features in the related condition resulted in higher activation values for those units, which in turn led to the faster growth of the connection weights between the features and the category label units (according to the CHL rule Equation 6, and as is shown in Equation 7), as compared with the unrelated condition. Importantly, a model such as Baywatch has no mechanism by which to account for the accelerated learning afforded by prior knowledge in the absence of preexisting concepts.

In both the related and the unrelated conditions, the frequency of the six features that were predictive of category membership varied. Whereas five of those features appeared frequently (with six or seven exemplars in each training block), the sixth appeared quite infrequently (one exemplar in each block). Murphy and Allopenna (1994) tested how subjects classified individual features during a test phase that followed learning, the results of which are presented in Figure 8. In the unrelated condition, RTs on single-feature classification trials were shorter for frequent than for infrequent features. In contrast, in the related condition, RTs were relatively insensitive to the features' empirical frequency. This pattern of results was also present in the subjects' categorization accuracy.

To determine whether KRES would also exhibit these effects, after training we tested the model on single features. The results are presented in Figure 8, superimposed on the empirical data. The figure indicates that KRES's RTs (as represented by the number of cycles the network needed to settle) reproduced the pattern of the human data. In KRES, infrequently presented related features were classified nearly as quickly as frequently presented ones, because during training, those features were activated by interfeature excitatory connections even on trials on which they were not presented. That explanation is documented in Figure 9A, which shows the average activation of category features during learning. In the related condition, infrequent related features were almost as active as frequent ones, with the result that connection weights between frequent and infrequent features and their correct category labels grew at almost the same rate (Figure 9B). The consequence was that the single-feature classification performance on the infrequent features was almost indistinguishable from that of the frequent features in the related condition (Figure 8). In contrast, in the unrelated condition, infrequent features were much less active, on average, than frequent ones, and hence, their connection weights grew more slowly. The consequence was that test performance on the infrequent features was much worse than on the frequent features in the unrelated condition.

As Figure 9B shows, at the end of training, the connection weights from frequent features were much larger in the unrelated condition than in the related condition, even though the subjects (and KRES) performed considerably better on the frequently presented related features than on the unrelated ones (a result seen in Simulation 3 as well). This result obtained because, during test, the single related feature activated all the other features to which it was related and all the related features together activated their category unit. In contrast, in the unrelated condition, the category unit received activation only from the single feature that was being tested. That is, the resonance among features in the related condition not

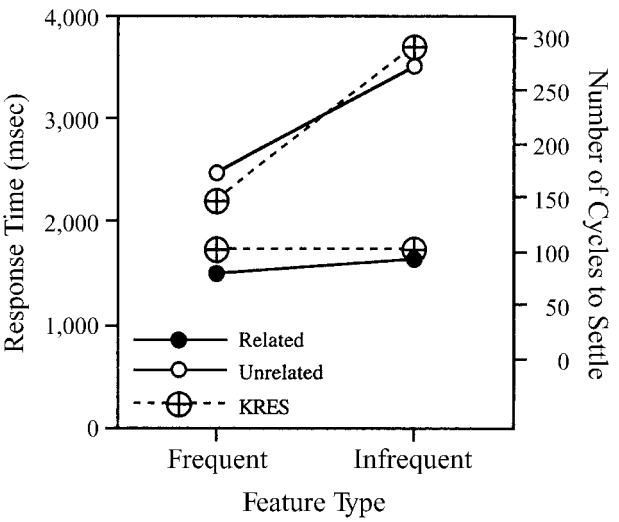

Figure 8. Results of single-feature tests in Murphy and Allopenna (1994), Experiment 2, and from Simulation 4.

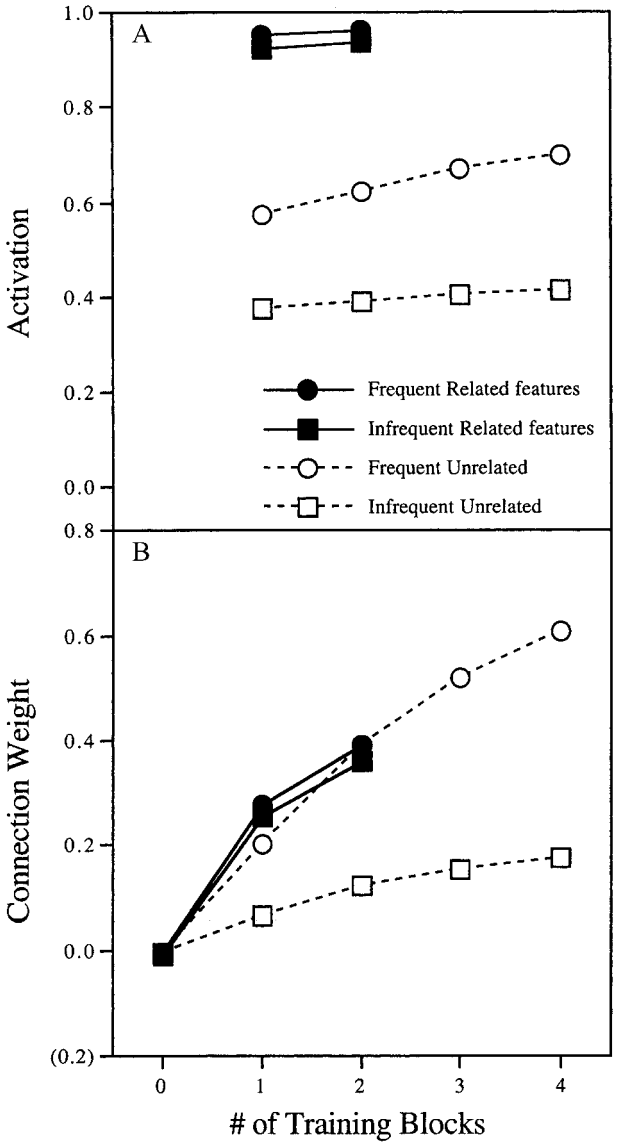

Figure 9. Average (A) activations and (B) weights to category label units in Simulation 4.

only helped during learning (by making the connection weights to grow more quickly), it also helped during test (by producing stronger activation of the category unit). As a result, the connections to the category units did not have to be as strong in the related condition as in the unrelated condition to achieve the same error rate, another reason why the error criterion was reached in fewer blocks in the unrelated condition.

The Separability of Prior Knowledge and Empirical Learning

The three previous simulations provide evidence in favor of KRES's ability to accelerate learning by introducing prior concepts (Simulations 2 and 3) and by amplifying the activation of features interconnected by prior knowledge via recurrent networks (Simulation 4). However, it can be shown that the success of these simulations did not depend on another distinctive characteristic of KRES—namely, that the output layer (i.e., the category label units) is recurrently connected to the features. Indeed, the empirical data we have considered thus far would also be consistent with a model in which only feature units (and perhaps prior concept units) were linked with recurrent connections. Once this constraint satisfaction network settled, activation could be sent to the output layer in a feedforward manner.

One reason why it is important to consider this alternative model carefully is that it is related to the question of whether the effects of knowledge and empirical learning can be conceived of as occurring independently—that is, in separate *modules*. For example, according to an *addition model* (Wisniewski & Medin, 1994), prior knowledge is used to infer new features, and those new features are input to the learning process alongside normal features. In addition, according to what Wisniewski and Medin called a *selection model,* prior knowledge selects (or weights) the features before they are input to the learning process. For both addition and selection models, knowledge and empirical learning can be considered separable, because knowledge works merely to transform the input that is provided to the empirical-learning module. In contrast, Wisniewski and Medin define a *tightly coupled* or *integrated* model for category learning as one in which prior knowledge and exemplars interact and, together, influence the learning process.

The KRES models used in Simulations 2 and 3 can be seen as examples of an addition model, because they introduced new *features* into the training pattern—what we have called *prior concepts* plus related features that were never presented. However, there are at least two ways that KRES implements integrated category learning. First, in Simulation 4, recurrent connections between feature units changed the effective weight of features by changing their activation values (because those changed activation values influenced the subsequent course of learning). This KRES model should not be seen as a mere selection model, however, because instead of a feature's *weight* being a fixed property of the feature, the feature activation values emerged dynamically as part of the resonance process. In other words, a feature's weight (i.e., its activation value) will vary depending on the set of features it appears with. Indeed, previous research has shown that the importance, or weight, of a feature will vary depending on the object in which it appears (Medin & Shoben, 1988).

KRES's assumption that activation flows not only forward from features to category labels, but also backward from category label units to features is a second way in which KRES implements an integrated model of category learning. That is, prior knowledge in the form of the connections emanating from the category label units affects the activation values of features, which in turn affects further learning. In the following two simulations, we present

evidence for this *top-down* effect of prior knowledge on empirical learning and, by so doing, provide additional evidence for a view of category learning that emphasizes the inseparable influences of knowledge and learning that occur during the acquisition of new categories.

Simulation 5: Learning Features Unrelated by Knowledge

Using a modified version of Murphy and Allopenna's (1994) materials, Kaplan and Murphy (2000, Experiment 4) provided a dramatic demonstration of the effect of prior knowledge on category learning. In that study, each category was associated with a number of knowledge-related and knowledge-unrelated features. However, the exemplars were constructed primarily from the latter: The training examples contained only one of the related features and up to five unrelated features that were predictive of category membership. The unrelated features formed a family resemblance structure much like that shown in Table 1. In contrast, because each exemplar had only one related feature, these features were related only to features in other exemplars. One might have predicted that the subjects would be unlikely to notice the relations among the related features in different exemplars, especially given that such features were surrounded by five unrelated features.

Kaplan and Murphy (2000) compared learning in this condition (the related condition) with learning in one that had the same empirical structure but no relations among features (the unrelated condition). In both conditions, there were features that were *characteristic* of the category because they appeared in so many category exemplars and also *idiosyncratic* features that appeared with just one exemplar. (These conditions were referred to as the *theme* and the *mixed-theme* conditions by Kaplan & Murphy, 2000.[5]) Kaplan and Murphy (2000) found that the subjects in the related condition reached a learning criterion in fewer blocks (2.67) than the unrelated group did (5.00). Thus, knowledge helped learning in the related condition despite the fact that there were very few feature relations that spanned category exemplars.

We simulated this experiment with a KRES model with 22 features on 11 binary dimensions. In the related condition only, the features within the two sets of 6 related features were interrelated with excitatory connections, as in Simulation 4. This represented the notion that these features were conceptually related prior to the experiment. The weight on these excitatory connections was set to 0.55, the weight on inhibitory connections was set to -2.0, the learning rate was set to 0.15, and the error criterion was set to 0.05. Each exemplar was constructed from 5 unrelated features and 1 knowledge-related feature, following Kaplan and Murphy's (2000) design. Given that each exemplar contained only 1 knowledge-related feature, it was unclear whether KRES would demonstrate an advantage for this condition over the unrelated condition that had no such prior knowledge.

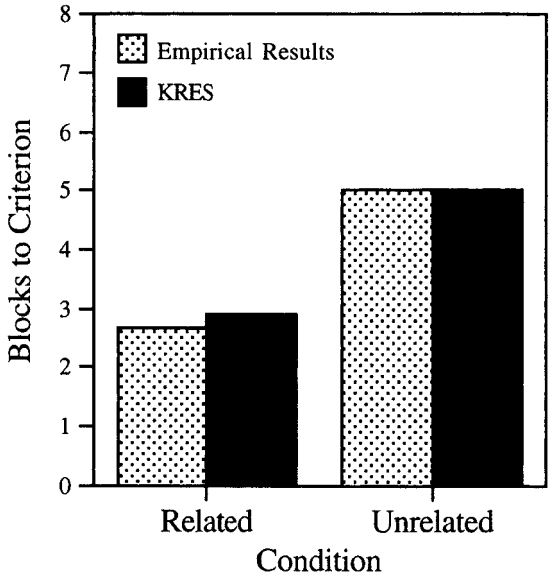

Figure 10. Results from Kaplan and Murphy (2000), Experiment 4, and from Simulation 5.

Figure 10 indicates that KRES did reproduce the learning advantage for the related condition, as compared with the unrelated condition, found with human subjects. This advantage obtained because even though each training example in the related condition contained only one knowledge-related feature, that feature tended to activate all the other features to which it was related, and hence the connections between the six related features and their correct category label were strengthened on every trial to at least some degree. That learning gave an advantage to the related group, which was identical to the unrelated group in terms of the statistical presentation of the exemplars and their features. For the unrelated group, the features that occurred only once per exemplar would be learned slowly, because of their low frequency. The resonance among those features in the related condition effectively raised their presentation frequency, thereby aiding learning.

In order to better understand what effect knowledge was having on the learning process, after training, Kaplan and Murphy (2000) presented test trials in which subjects were required to perform speeded classification on each of the 22 features. Figure 11 presents the results of these tests, indicating that the subjects in the unrelated condition were faster at classifying those features that appeared in several training exemplars (characteristic features) than those that appeared in just one training exemplar (idiosyncratic features). In contrast, in the related condition, the subjects were faster at classifying the idiosyncratic features, which for them were related features. Importantly, the subjects in the related condition were no slower than the unrelated subjects at classifying the characteristic features (i.e., the unrelated features), even

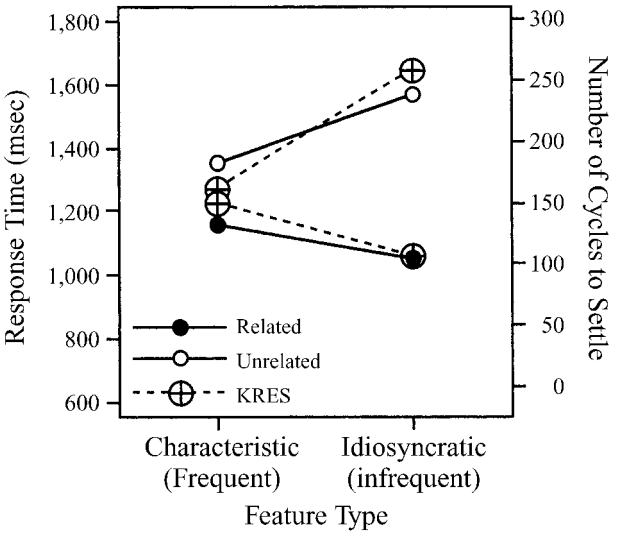

Figure 11. Results of single-feature tests in Kaplan and Murphy (2000), Experiment 4, and from Simulation 5.

though those features were not related to the other features and even though they had experienced fewer training blocks on average (2.67 vs. 5.00). That is, the prior knowledge benefited the features related to knowledge but did not interfere with the features that were not related to it.

The latter result is a challenge for many standard connectionist accounts of learning, because, as we saw in Simulation 1, in such accounts the better learning associated with related features would be expected to compete with and, hence, overshadow the learning of unrelated features. In contrast, Figure 11 indicates that KRES is able to account for the better learning of the related features (the related-condition–idiosyncratic features in the figure) without entailing a problem in learning unrelated features (the related-condition–characteristic ones). This result can be directly attributed to the use of recurrent connections to the category label units. After some excitatory connections between the characteristic features and the category labels have been formed, the subsequent presentation of these unrelated features activates a category label, which in turn activates the associated related features, which in turn activate one another, which in turn increase the activation of the category label and then the unrelated features. This greater activation of the unrelated features leads to accelerated learning of the connection weights between the unrelated features and the category labels.

These results indicate that when there are existing category features to which new features can be integrated, KRES's recurrent network that allows activation to flow from category labels to features can compensate for the effects of cue competition. Indeed, Kaplan and Murphy (2000) presented evidence suggesting that the better learning of unrelated features in the related condition arose in part from subjects' integrating those features with the other features. KRES provides a potential mechanism by which such integration is carried out: Unrelated features become linked to the related ones indirectly through the category labels. Although it is likely that the subjects' integration processes often involved more complex explanatory reasoning (e.g., inferring a reason for why arctic vehicles should have air bags rather than automatic seat belts), the indirect connections between unrelated and related features formed by KRES may be a necessary precondition for such reasoning.

We should point out that the question of exactly when and how much knowledge helps the learning of knowledge-unrelated features is a delicate one, because sometimes knowledge-unrelated features are learned better in the related condition (the one Kaplan & Murphy, 2000, simulated, although this effect was not significant) and sometimes the two do not differ (e.g., Kaplan & Murphy, 2000, Experiment 5). This effect probably depends on a number of factors, including the degree to which the knowledge-related and knowledge-unrelated features can themselves be related, the statistical category structure, and various learning parameters (see Kaplan & Murphy, 2000, for a discussion). However, the main point is that, counter to the prediction of most error-driven learning networks, knowledge does not hurt the learning of unrelated features, and KRES is able to account for this effect or even for an advantage, when one occurs.

Finally, KRES's success at accounting for classification performance in the unrelated condition in this simulation, as well as in the previous one, is notable, because the difference in classification performance on the frequent and the infrequent features in Simulation 4 and between characteristic and idiosyncratic features in Simulation 5 are examples of *feature frequency effects* in which features are more strongly associated with a category to the extent to which they are observed in more category exemplars (Rosch & Mervis, 1975). Again, this result demonstrates that KRES can account for knowledge advantages and more data-driven variables within the same architecture. With prior knowledge (excitatory internode connections), KRES exhibits the accelerated learning and the resulting pattern of single-feature feature classifications found in the empirical studies presented in Simulations 2–5. Without that knowledge (i.e., without those connections) KRES reverts to an empirical-learning model that exhibits standard learning phenomena, such as a prototype advantage and cue competition (Simulation 1) and feature frequency effects (control conditions of Simulations 4 and 5).

Revising Prior Knowledge

In the simulation of knowledge effects presented so far, we have allowed KRES to learn new connections to category label units, but we have disabled learning on those connections that represented prior knowledge. Our reason for doing so was based on the belief that in many cases (and specifically, in the situations modeled in Simulations 2–5), prior knowledge is highly entrenched and,

hence, is unlikely to be greatly altered in a category-learning task. For example, it would be difficult to get subjects to change their minds about how wings enable flying or whether arctic vehicles need protection from the cold in the course of a brief category-learning experiment. However, there might be other cases in which subjects have little at stake in the knowledge they apply to a learning situation and so might be willing to update that knowledge in light of empirical feedback. It seems quite reasonable, or perhaps necessary, therefore, to make a distinction between knowledge that is likely versus unlikely to be changeable by experience of this sort.

In our final simulation, we demonstrate the ability of CHL to revise nonentrenched prior knowledge. We examined how the CHL rule updates weights on connections involving not only category label units, but any connection in the network, including those that represent prior knowledge. We considered a case in which the prior knowledge in question involved the interpretation of novel perceptual stimuli. As the empirical results will show, the subjects in this experiment apparently were not strongly committed to how they initially interpreted these stimuli and, hence, were amenable to changing their interpretation in light of feedback.

Our expectation was that the CHL rule would change connection weights in a manner consistent with incoming empirical information. Indeed, we had run versions of all four of the previous simulations in which we allowed the prior knowledge connections to be changed. Generally speaking, the connections tended to become stronger—that is, negative connections became more negative, and positive connections became more positive. This result was expected, because the empirical structures of the training stimuli were consistent with the prior knowledge. In contrast, in Simulation 6, empirical feedback was inconsistent with some of that knowledge, and we expected that prior knowledge would get weaker as a result.

A second purpose of Simulation 6 was to present more evidence for the claim that activation flows not only forward from features (and perhaps prior concepts) to category labels, but also back from the category labels. We will show that how one interprets novel perceptual stimuli depends on their possible categorizations. That is, top-down knowledge, in the form of already known category labels connected with prior knowledge, can influence how one interprets unfamiliar stimuli.

Simulation 6: Interpreting Ambiguous Stimuli and Updating Prior Knowledge

Wisniewski and Medin (1994, Experiment 2) showed subjects two categories of drawings of people that were described as drawn by creative and noncreative children or by farm and city kids. Wisniewski and Medin used line drawings to illustrate that what constitutes a feature in a stimulus depends on the prior expectations that one has about its possible category membership. For example, they found that subjects assumed the presence of *abstract features* of a category on the basis of the category's label (e.g., they expected creative children's drawings to depict unusual amounts of detail and characters performing actions). The subjects examined the drawings for concrete evidence of those expected abstract features and, as a result, noticed different features, depending on their expectations. Moreover, Wisniewski and Medin found that the feedback that the learners received about category membership led them to change their original interpretation of certain features of the line drawings. For example, after first interpreting a character's clothing as a farm uniform (and categorizing the picture as having been drawn by a farm kid), some subjects reinterpreted the clothing as a city uniform after receiving feedback that the picture had been drawn by a city kid.

To fully account for these effects with KRES would require a much more detailed perceptual representation scheme and, perhaps, a more sophisticated inference engine. However, it is also possible that the resonance process we have described could account for some of these reinterpretation effects. The basic requirements are that category feedback be able to influence lower level connections between perceptual properties and their interpretation and that the relevant prior knowledge not be too entrenched, so that interpretations can be altered. (Presumably, it would have been difficult for Wisniewski & Medin's, 1994, subjects to learn to interpret long hair as being short or to adopt other interpretations that grossly flouted past experience.)

To demonstrate these effects with KRES, we imagined a simplified version of the materials in Wisniewski and Medin (1994), in which there were only two drawings. One drawing (Drawing A), was of a character performing an action interpretable either as climbing in a playground or dancing. (Two of their subjects actually gave these different interpretations of a single picture; see p. 260.) This drawing demonstrated how ambiguous input could be interpreted on the basis of category information. In the other drawing (Drawing C), a character's clothing could be seen as a farm uniform or a city uniform. These alternative interpretations are represented in the left side of the KRES model in Figure 12. Whereas we assumed that the two interpretations of Drawing A were equally likely, we assumed that a city uniform would be the more likely interpretation of Drawing C (as depicted by the heavier line connecting the features of Drawing C and the city uniform interpretation). This example will demonstrate how incorrect expectations can be unlearned. The alternative interpretations were connected with inhibitory connections representing that only one interpretation was correct: The clothing could not be both city and farm garb.

In a more complete simulation of this process, the perceptual features at the left of Figure 12 would be more lawfully related to different interpretations. For example, some aspects of a picture would suggest dancing, and an overlapping set would suggest climbing. In this simplified version, we simply associated the entire set with the

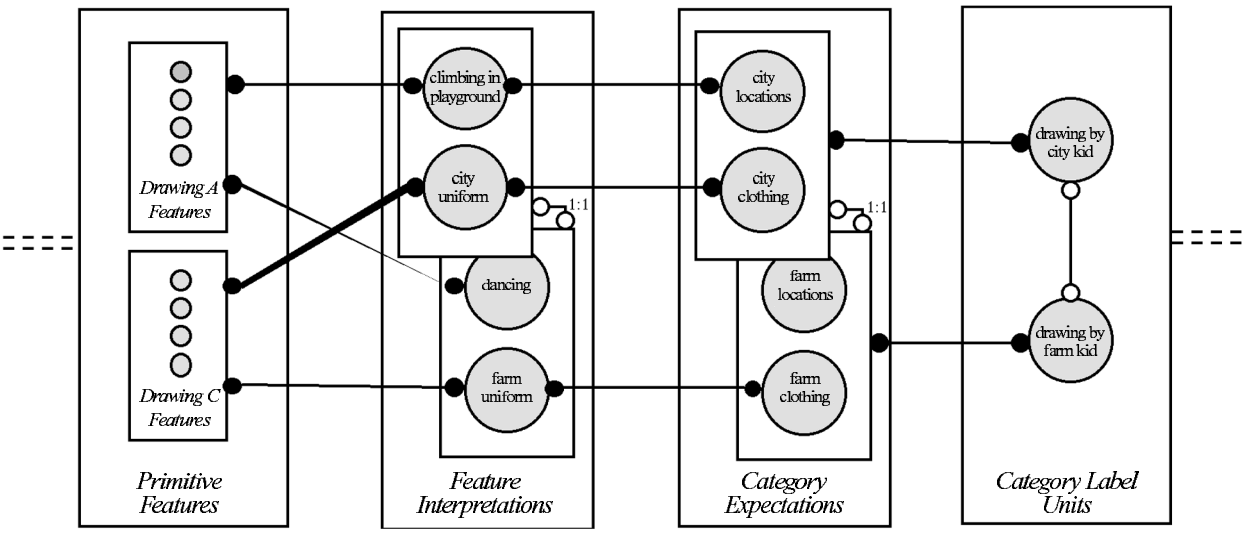

Figure 12. KRES model for Wisniewski and Medin (1994, Experiment 2).

picture's possible interpretations. The assumption underlying the model is that there are intermediate descriptions of the primitive features that intervene between the sensory processes and category information. However, as considerable recent research has shown (Goldstone, 1994; Schyns & Murphy, 1994; Schyns & Rodet, 1997), the interpretation of perceptual primitives can change as a result of experience in general and of category learning in particular.

The model in Figure 12 was presented with the problem of learning to classify Drawing A as done by a city kid and Drawing C by a farm kid. We represented the expectations that learners form in the presence of meaningful category labels, such as farm or city kids, as units connected via excitatory connections to the category labels, as shown in the right side of Figure 12. The model expected city and farm kids to be in locations and to wear clothing appropriate to cities and farms, respectively. These expectations were, in turn, related by excitatory connections to the picture interpretations that instantiated them: Climbing in a playground instantiated a city location, and city and farm uniforms instantiated city and farm clothing, respectively. Finally, because people know what climbing children look like and have some idea about the appearances of city and farm clothes, these interpretations were, in turn, associated with perceptual features. All the inhibitory connections shown in Figure 12 were set to -3.0, and all excitatory connections were set to 0.25, except for those between Drawing C's features and their city uniform interpretation, which were set to 0.30.

Before a single training trial was conducted, KRES was able to decide on a classification for both drawings on the basis of its prior knowledge. Upon presentation of Drawing A, its two interpretations—climbing in a playground or dancing—were activated, and climbing in a playground, in turn, activated the city location node, which in turn activated the category label for city kids' drawings. The drawing was correctly classified as having been drawn by a city kid. Moreover, as the network continued to settle, activation was sent back from the category label to the climbing in a playground unit. As a result, that interpretation of Drawing A was more active than the dancing interpretation when the network settled. Because dancing was not associated with either of the relevant categories, this interpretation of the drawing was deemphasized, even though perceptually it was just as consistent with the input. That is, the top-down knowledge provided to the network (the category labels and their associated properties) resulted in the resolution of an ambiguous feature. Wisniewski and Medin (1994) found that the same drawing would be interpreted as depicting dancing instead when subjects were required to classify the drawings as having been done by creative or noncreative children.

What happens when the model's expectations are incorrect? One potential problem with models that use prior knowledge is that their knowledge may overwhelm the input, so that they hallucinate properties that are not there. Any such model must be flexible enough to use knowledge when it is appropriate but also to discover when it is incorrect for a given task. Upon presentation of Drawing C, its two interpretations were activated, but because the city uniform interpretation received more input as a result of its larger connection weight, it quickly dominated the farm uniform interpretation. As a result, the category label for city kids' drawings became active (via the city clothing expectation). However, recall that this drawing was, in fact, made by a farm kid, and so this categorization was incorrect. This mistake generated error feedback, which in turn resulted in a change of the drawing interpretation. KRES did this be-

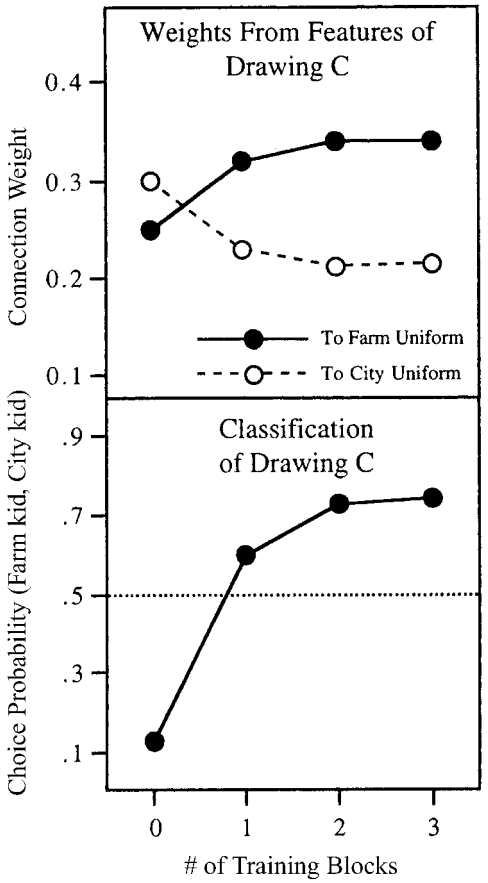

Figure 13. Connection weights and classification results from Simulation 6 as a function of the number of blocks of training.

cause, during the model's plus phase, the farm kids' category label was more active than the city kids' label, as a result of the external inputs to those units. The activation emanating from the farm kids' label led to the activation of the farm clothing expectation and then the farm uniform feature interpretation, which ended up dominating the city uniform unit.

This result indicates that KRES can reinterpret features in light of error feedback. The more important question, however, is whether KRES can *learn* this new interpretation so that Picture C (or a similar picture) will be correctly classified in the future. The top panel of Figure 13 shows the changes to the connection weights brought about by the CHL rule, with a learning rate of .30 as a function of the number of blocks of training on the two drawings. This figure indicates that the connection weights associated with the interpretation of Drawing C as a city uniform rapidly decreased from their starting value of .30, whereas the weights associated with the farm uniform interpretation increased from their starting value of .25. As a result, after just one training block, KRES's classification of Drawing C switched from having been done by a city kid to having been done by a farm kid (as indicated by the choice probabilities shown in the bottom panel of Figure 13). KRES uses error feedback to learn a new interpretation of an ambiguous drawing, just as human subjects do (Wisniewski & Medin, 1994).

This version of KRES illustrates the importance of distinguishing between the fairly raw input and the interpretation of that input (although the interpretation involves the grouping of perceptual features that may itself have perceptual consequences, as in Goldstone, 1994, and Goldstone & Steyvers, 2001). If the drawings were considered to be single-input units, this learning would not be possible; or if there were no interpretation units, the meaning of the features could not be learned—only the pattern's ultimate categorization. This learning is important, however, because it can then be applied to new stimuli. If a picture with some of the same perceptual units were presented after this learning phase, its interpretation would also be influenced by the interpretations of Drawings A and C. Thus, distinguishing interpretations from the input features, on the one hand, and category units, on the other, allows KRES to use knowledge to flexibly perceive input. One might worry about the use of category feedback to greatly change perceptual structure. However, extremely well entrenched perceptual generalizations would presumably not be unlearned as the result of learning a single new category.

A central point about this simulation is that it reveals how experience can affect knowledge and vice versa. On the one hand, prior knowledge about the categories influenced the perceptual interpretation of the ambiguous pictures. On the other hand, experience (in the form of feedback) with the farm kid's drawing changed the model's prior expectation about what a city uniform would look like. That is, background knowledge not only influences category learning, category learning influences one's knowledge. Capturing the interplay between learning and knowledge is one of the main goals of KRES.

Of course, how much knowledge is affected by feedback will depend on how committed the learner is to that knowledge. For Wisniewski and Medin's (1994) subjects, nothing much depended on their beliefs about how farm uniforms look or how much detail is in the drawings of creative children, nor did they have much prior experience with these categories. This is exactly the sort of knowledge that would be flexible in the face of evidence.

GENERAL DISCUSSION

We have presented a new model of category learning with which we attempt to account for the influence of prior knowledge that people often bring to the task of learning a new category. Unlike past connectionist models of category learning that have used feedforward networks, KRES uses a recurrent network in which prior knowledge is encoded as connections among units. We have shown that the changes brought about by this re-

currently connected knowledge provide a reasonable account of five empirical data sets exhibiting the effects of prior knowledge on category learning.

We have taken pains to be clear on which of the distinctive characteristics of KRES were responsible for the success of the various simulations. In Simulations 2, 4, and 5, we demonstrated how KRES's recurrent network provides a pattern of activation among units that would account for the finding that prior knowledge accelerated the learning of connections to the label of a new category. In Simulation 2, we demonstrated such accelerated learning when the features of a category activated a common preexisting concept. In Simulations 4 and 5, accelerated learning was demonstrated when category features were related only to one another. We also showed how prior knowledge connections among features led to them being classified correctly on a transfer test even when they had been presented during training with low frequency (Simulations 4 and 5) or not at all (Simulation 3).

Simulations 3–5 demonstrated that both people and KRES exhibit considerable learning of features not related by prior knowledge. Indeed, the results of Simulation 5 indicate that knowledge can aid (or at least not hurt) the learning of unrelated features, a striking result in light of the well-known learning phenomenon of overshadowing. KRES's success at simulating this result provides an important piece of evidence for our claim that activation can flow backward from category labels to features, a natural consequence of KRES's use of recurrent networks. The top-down flow of activation was also instrumental in Simulation 6, in which excitatory connections from meaningful category labels resolved the ambiguous interpretation of a feature.

The final important property of KRES is its use of a CHL rule. This rule allows the learning of connections not directly connected to the output layer, including the *unlearning* of knowledge that is inappropriate for a particular category. Simulation 6 demonstrated how the knowledge that led to one interpretation of an ambiguous feature could be unlearned and a new interpretation learned when the network was provided with feedback regarding the stimulus's correct category.

In the section that follows, we will discuss the interactions between knowledge and data during category learning that are accounted for by KRES. We then will discuss some of KRES's inadequacies as an empirical-learning system and some possible solutions to those problems. We next will discuss possible extensions to KRES regarding the representation of knowledge and will consider the ultimate source of that knowledge. Finally, we will discuss the use of recurrent networks in KRES and other cognitive models.

The Interaction of Knowledge and Data in Category Learning

There have been very few attempts to account for the effects of both prior knowledge and empirical information on category learning in an integrated way. As was discussed earlier, many researchers in the field seem to have adopted a divide-and-conquer approach in which they assumed that the effects of knowledge and empirical learning can be studied independently and have focused on the empirical-learning part (often considered the *basic learning* component). The role of knowledge is often limited to the selecting or weighting of features (a selection model) or to inferring new features (an addition model), which are then input into the basic learning module—examples of what Wisniewski (1995) has called the *knowledge-first* approach to category learning. Alternatively (or in addition), knowledge might come into play after empirical regularities have been noticed, an example of an *empirical-first* approach. In either approach, prior knowledge and empirical learning are considered to be separate modules, an assumption that licenses the study of one in isolation from the other.

Wisniewski and Medin (1994; Wisniewski, 1995) and Murphy (2002) have criticized the view that knowledge and empirical learning can be treated as separate modules in this way. The rationale for independent modules can apply only if knowledge effects do not interact with the basic learning process or, for that matter, with other processes that involve concepts, such as induction, language processing, categorization, and so on. If these processes do interact with prior knowledge, the modular approach may be not just incomplete but incorrect for a real-world case in which learners have some prior knowledge about the domain. For these reasons, Wisniewski and Medin argued for an integrated model of concept learning that acknowledges the interacting influences of knowledge and empirical information.

There are several ways in which KRES exemplifies this sort of integrated learning. First, in Simulation 4, we showed how knowledge in the form of recurrent connections among feature units changed the activation values of those units, which in turn influenced the learning process. Because these activation values are determined by the constraint satisfaction process, in KRES the importance of a feature for learning depends on the set of features it appears with, rather than on its being context independent (Medin & Shoben, 1988). Second, in Simulations 5 and 6, we showed how recurrent connection from category labels also influenced learning. In particular, in Simulation 6, we demonstrated how top-down knowledge influenced the features that were "observed" in ambiguous stimuli. Finally, in Simulation 6, we also showed how empirical information in the form of error-correcting feedback permanently changed that knowledge in such a way that different features were observed in the same stimuli. These mutual influences of knowledge on data and vice versa are just some of those that motivated a call for an integrated account of learning (Wisniewski & Medin, 1994).

While emphasizing KRES's integrated approach to category learning, we have also stressed that KRES accounts for many aspects of normal empirical learning as well. For example, in Simulations 3–5, we demonstrated

how KRES exhibits learning of features not related by prior knowledge even when they appear alongside related features. In Simulation 1, we showed how, in the absence of any prior knowledge, KRES exhibits typicality and cue competition effects, and in the control conditions of Simulations 4 and 5, we showed KRES exhibiting feature frequency effects. In other words, KRES exhibits interactions between knowledge and data when knowledge is present, but when it is not, KRES reverts to an empirical-learning model that exhibits some of the standard phenomena of associative learning.

In this light, we believe that KRES offers a new perspective on the nature of the interaction between prior knowledge and empirically based learning processes. In the KRES architecture, knowledge can be added on to a model with no prior knowledge in the form of preexisting concepts and connections. However, when it is added on, it may interact quite strongly with incoming empirical information, producing, as a result, the kinds of dramatic effects on learning performance seen in humans. KRES exhibits these qualities because it possesses the nonlinear activation dynamics (recurrent networks) that result in the (nonlinear) effects on behavior that have been taken as evidence for the inseparability of knowledge-driven and empirical-driven learning. The result, we suggest, is a model that offers a framework in which to pursue issues in knowledge-based learning, experience-based learning, and the interaction between the two.

We believe that a unified approach to empirical and knowledge-related learning is necessary because people's knowledge of most real-world categories involves a blend of the two types of information. Even when real-world category learning is mostly determined by empirical input, cases in which learners have no prior knowledge that links features to prior concepts and each other are rare. And even when learning is dominated by a learner's prior theory, we believe, like Keil (1995), that "all theories run dry" eventually and that the category will exhibit features and interfeature correlations that are unexplained by the theory. Because people's knowledge of most categories includes both theoretical and empirical information, it is important for a model of category learning to accommodate both.

KRES as a Model of Empirical Learning

We have stated our commitment to a unified approach to empirical and knowledge-based learning and have noted KRES's strengths as an empirical-learning system. However, it is also important to note its weaknesses. One important limitation of KRES as currently formulated is that it is unable to solve nonlinearly separable categorization problems in the absence of prior knowledge. Nonlinearly separable problems are those such as XOR or, more generally, cases in which a category cannot be summarized by a single central tendency. For example, learning the concept of *birds* would be a nonlinearly separable problem if one thinks of penguins as being more similar to seals and otters than they are to cardinals and chickens. People's ability to learn some nonlinearly separable categories has been taken as an important piece of evidence in favor of exemplar models of concepts (Medin & Schwanenflugel, 1981).

Another deficiency is that KRES has no representation of the importance, or weight, of individual dimensions of the stimulus space on categorization judgments. In comparison, existing similarity-based models (including exemplar models) account for the fact that classifiers learn to optimally allocate attention, when classifying, by incorporating per-dimension attention weights (Kruschke, 1992; Nosofsky, 1984; Rosch & Mervis, 1975). More recent models also implement a limited-capacity attention and specify how attention changes (and how those changes are learned) with error feedback (Kruschke, 1996a, 1996b, 2001; Kruschke & Blair, 2000; Kruschke & Johansen, 1999). There has also been a renewed emphasis on the importance of rule-based classification learning (Nosofsky, Palmeri, & McKinley, 1994) and on specifying how rules interact with exemplars (Erickson & Kruschke, 1998; Smith, Patalano, & Jonides, 1998), or an implicit learning system (Ashby, Alfonso-Reese, Turken, & Waldron, 1998; Ashby & Waldron, 1999).

One reason we did not begin our efforts with a model that already accounts for one or more of these empirical-learning phenomena is that it is unclear how such effects manifest themselves when prior knowledge is present. For example, much research has shown that there is no advantage for linearly separable or nonlinearly separable categories, a result consistent with exemplar theories. However, Wattenmaker et al. (1986) found that they could create an advantage for either kind of category depending on the subjects' knowledge structures. Similarly, Murphy and Kaplan (2000) found that when categories had a unifying theme (such as the arctic and jungle vehicles of Simulation 4), the nonlinearly separable categories became difficult to learn. In other words, what may be important is not the category's empirical structure per se (e.g., linearly separable or nonlinearly separable) but, rather, whether that structure matches or mismatches the expectations induced by the learner's knowledge (Waldmann, Holyoak, & Fratianne, 1995). How learners allocate attention to features and generate candidate classification rules is also likely to interact heavily with their domain knowledge (e.g., Pazzani, 1991).

That said, it is also clear that people do remember individual cases (e.g., the neighbor's dog, one's own car), have attention limitations, and generate categorization rules, and in future work we intend to upgrade KRES to address more of these empirical-learning phenomena.[6] However, our point is that many of these effects are likely to interact strongly with knowledge, and KRES is well placed to address such interactions, rather than just the effects as they arise in the perhaps unusual situation in which features are meaningless and have no prior relations.

KRES and the Representation of Prior Knowledge

In KRES, we have looked at two forms of prior knowledge that might be involved in category learning. First, we followed the lead of Heit and Bott (2000) by assuming the presence of a prior concept that was similar to the to-be-learned category. The prior concept helped learning by itself being associated to the new concept name. Second, we assumed the presence of connections among features. These connections sped learning by increasing the feature units' activation values and, thus, the rate of growth of the connections to the new concept.

Although these two simple forms of knowledge were sufficient to account for the empirical results we simulated, we do not claim that all forms of knowledge can be represented with just simple associations among concepts. For example, many theorists have argued for the importance of representing knowledge in the form of propositions, a capacity that entails the need to bind concepts to their roles as arguments of predicates (Fodor & Pylyshyn, 1988; Holyoak, 1991; Marcus, 2001). On the one hand, our decision to not include more structured representations in KRES was partly pragmatic: The question of how best to accomplish variable binding in connectionist-style networks remains open (cf. Holyoak & Thagard, 1989; Hummel & Holyoak, 1997; Shastri & Ajjanagadde, 1993; Smolensky, 1990; Thagard, 1989; Touretzky & Hinton, 1988). On the other hand, we believe that the success of the simulations reported here suggests that our simple approach to knowledge representation may be capturing much of what is essential about how knowledge affects category learning. For example, symmetric interconnections might be sufficient to model the effects of a number of different types of semantic relations, including causal relations (wings enable flying, flying enables an animal to roost in trees, etc.), feature co-occurrence (small birds tend to sing, large birds do not), function–form relationships (animals with big eyes see well at night), or generalizations across a large domain (baby animals are smaller than the adults). That is, although knowledge is often more structured than interconcept associations, simpler forms of knowledge representation might turn out to be adequate for modeling many phenomena.[7]

Another model that has been proposed to account for the effects of prior knowledge on category learning is the integration model (Heit, 1994, 1998). The integration model builds on existing exemplar models by assuming that prior knowledge takes the form of exemplars from other, already learned categories. According to this account, knowledge of, say, the causal connection between wings and flying would consist of no more than the co-occurrence of wings and flying in one's memory of animals that one has observed in the past. Using this exemplar-based representation of knowledge, Heit (2001) simulated the effects of a number of studies in which knowledge effects have been looked at, including some of the ones we have examined here.

The possibility that prior knowledge might be represented as nothing more than previous experiences is an important one because, if true, it implies that there is no need for inductive processes that abstract such facts as *wings cause flying*. In comparison, by representing the relation between wings and flying as an excitatory connection, KRES assumes the existence of knowledge that is generic, or abstract, because it is independent of any particular context (i.e., exemplars). There are several reasons why we have chosen to represent knowledge in the form of abstract interconcept relations in KRES, rather than as previously observed exemplars. First, we believe that many knowledge effects in category learning arise from facts about whole classes of objects, rather than about particular exemplars. For example, people know a wide range of facts about animals—that animals need food and shelter to survive, that animals are of the same species as their parents, and that animals with wings usually fly (and that wings support the animal's body on the air, that flying is a useful evolutionary advantage, etc.). But these are facts that one holds to be true about animals in general, not ones stored with individual category members. For example, although while learning about a new species of songbird, one might be reminded of the robin that one sees in the backyard every morning, it is not that familiar robin that leads to one's expectations about the new bird's need for nourishment, its parentage, or its evolutionary history.

Second, even if prior exemplars were a source of general knowledge, prior research suggests that the only ones likely to be retrieved are those that are highly similar to exemplars of the new category. That is, one's backyard robin is unlikely to come to mind while one is learning about dissimilar birds, such as eagles, penguins, or ostriches (to say nothing of bats or flying squirrels). Indeed, it was because Murphy and Allopenna (1994) taught people such concepts as jungle and arctic vehicles that did not generally remind them of known categories that we chose not to represent prior knowledge in the form of prior concepts in Simulation 4 as we did in Simulations 2 and 3. In contrast, in his own simulation of Murphy and Allopenna's results, Heit (2001) argued that "Although participants may never have seen a vehicle with all of the characteristics of the Integrated [Related] prototype, they probably knew of real objects that preserve some of the predictive relations between the features. For example, they might have had prior examples of lightly insulated jungle buildings, green clothing in jungles, jungles in Africa" (p. 168). However, because the assumption of exemplar models is that only highly similar exemplars are usually retrieved or have much influence on the categorization process (Nosofsky & Palmeri, 1997), it seems very unlikely that prior exemplars of clothing or buildings came to mind when people were learning about the new type of vehicle. Indeed, previous research has shown that knowledge embedded in previous examples is unlikely to transfer across disparate domains until it has been abstracted from its original

context (Catrambone & Holyoak, 1989; Ross & Kennedy, 1990). Rather than assuming that learners' prior knowledge is limited to only those highly similar exemplars that they happen to be reminded of, KRES assumes the availability of the full range of world knowledge, including (in the case of Murphy and Allopenna's subjects), such facts as that jungles are hot, that insulation retains heat, that ice is slippery, requiring some form of traction, and so on.[8]

Finally, even when learners are reminded of exemplars from existing categories, the integration model does not specify the mechanisms by which those exemplars aid learning. For example, in Heit's (2001) simulation of Murphy and Allopenna (1994), the prior exemplars were already associated with the new categories. This assumption is clearly unrealistic—subjects had no way of knowing at the start of the experiment that the jungle vehicles would be called DAX and the arctic vehicles would be called KEZ. In contrast, such models as Baywatch (Heit & Bott, 2000) and KRES specify how new associations to the unfamiliar category labels are learned and how that learning is accelerated by prior concepts (in the case of Baywatch and KRES) or by the presence of prior interfeature relations (in the case of KRES). That is, a mechanism by which prior exemplars influence category learning remains to be specified.[9]

Further research will be needed to determine the relative contributions to category learning of knowledge that is abstract or generic (birds' parentage and evolutionary history) and knowledge encoded in the form of previously observed cases or exemplars (the robin in the backyard). We have given our reasons for incorporating abstract knowledge in KRES, but we acknowledge that prior exemplars may turn out to be important in special cases, as when they are highly similar to the new category being learned. For those situations in which prior exemplars turn out to be important, the KRES architecture can be easily upgraded to include them (see note 9).

Where Does Knowledge Come From?

Our emphasis on KRES's use of generic prior knowledge of course leaves open the question of where that knowledge comes from in the first place. On the one hand, we believe that much of this knowledge is acquired through explicit instruction or is generated by learners' own inferential processes, and we consider it an advantage of the KRES architecture that it can accommodate these sources. However, we also believe that abstract knowledge often derives from direct experience, and the nature of the inductive processes that generate this knowledge is an open question of considerable theoretical interest. We believe that KRES itself may provide some insight into this question. First, the same kinds of processes as those involved in category learning could result in the learning of associations between features. For example, noticing that wings and flying covary could be incipient knowledge of aeronautics and could influence learning about flying animals. We did not invoke such a process, because we were primarily simulating experiments that relied on previously known, well-entrenched knowledge. But the CHL algorithm could be used for associative learning of feature links as well. Indeed, Simulation 6 illustrated that CHL could *revise* prior knowledge when it was inconsistent with error feedback. With enough such experience, a model could permanently learn that this knowledge was incorrect.

Another important kind of inductive learning that KRES may be able to accomplish is the learning of feature vocabulary. For example, although we described Simulation 6 as KRES's reinterpreting an existing feature set, it may be equally valid to consider that simulation to be a case of learning a new feature vocabulary, one that was more useful for the learning task at hand (Goldstone, 2000; Goldstone & Steyvers, 2001; Schyns & Rodet, 1997). Although Simulation 6 did not implement this process (because the different interpretations were already related to the perceptual units before the experiment started), we believe that KRES is one way to start addressing this claim computationally. If sensory or perceptual units are thought of as being grouped to form higher level units, experience in the form of top-down error feedback will likely influence that grouping (for a related approach, see Goldstone, Steyvers, Spencer-Smith, & Kersten, 2000).

Recurrent Networks and Cognitive Models

KRES is, of course, not the first cognitive model to make use of recurrent networks. One early example is the interactive activation and competition (IAC) model of word perception (McClelland & Rumelhart, 1981; Rumelhart & McClelland, 1982). Like KRES, IAC uses the spread of activation from higher to lower level nodes to incorporate the effects of top-down knowledge in cleaning up and identifying input patterns (i.e., letters). Constraint satisfaction networks have also been used to model higher level cognitive phenomena, such as analogy (Holyoak & Thagard, 1989), explanation (Thagard, 1989), and decision making (Holyoak & Simon, 1999; Thagard & Millgram, 1995) and a variety of phenomena in the social psychology literature (Kunda & Thagard, 1996; Read & Miller, 1994; Shultz & Lepper, 1996; Spellman & Holyoak, 1993). However, unlike KRES, these models have not addressed the issue of learning that has been so central in the empirical studies we have simulated here.

One category-learning model that uses a recurrent network is Goldstone's (1996) RECON. Because RECON's purpose was to account for certain category-learning effects unrelated to prior knowledge (the effect of nondiagnostic features and the caricature effect), it does not represent knowledge (although it presumably could do so in the same manner as KRES). A more important difference is that RECON's Hebbian learning algorithm is insensitive to whether a classification error is committed. In contrast, CHL reflects the error-driven nature of associative learning in both animals and humans.

Recurrent networks that use an error-driven learning algorithm are common in the domain of language processing, including models of word recognition and lexical processing (e.g., Hinton & Shallice, 1991; McLeod, Shallice, & Plaut, 2000; Plaut & Booth, 2000), speech perception (e.g., Gaskell & Marslen-Wilson, 1997), speech production (e.g., Dell, Juliano, & Govindjee, 1993), and sentence comprehension (e.g., Christiansen & Chater, 1999; Tabor, Cornell, & Tanenhaus, 1997). In these language-processing models, it is common to employ versions of backpropagation suitable for recurrent networks (Almeida, 1987; Pearlmutter, 1995; Pineda, 1987), instead of the CHL rule that we have used. Our choice of CHL was motivated by claims of its greater biological plausibility and faster learning relative to backpropagation (O'Reilly, 1996). However, our demonstration of the equivalence of CHL to the delta rule in Simulation 1 under certain circumstances indicates that it may be relatively difficult to distinguish between these learning rules on the basis of behavioral data alone. At least regarding the empirical studies we have simulated here, we have no reason to believe that a recurrent version of backpropagation would not have fared as well as CHL.

Although the ability to naturally represent prior knowledge in the form of excitatory and inhibitory connections among concepts is an important advantage of recurrent networks, it raises the question of how the strengths of those connections should be chosen. Indeed, if one were to count the strength of each preexisting connection as a free parameter, these models can be seen as having a large number of parameters, leading to the standard problem of data overfitting. In the present work, this problem was addressed by constraining each simulation to have one strength value (two in Simulation 6) for all excitatory connections and another for all inhibitory connections. As a result, each model fit was achieved by adjusting only a relatively small number of free parameters: the excitatory and inhibitory connection strengths, the learning rate, and the error criterion.

We expect that computer models of prior knowledge effects on category learning will evolve quickly in the future and that, as they do, the well-known methods of quantitative model fitting and model comparison will be called upon in order to discriminate among competing theories. In the present simulations, our goal was only to provide a good qualitative account of the empirical phenomena, and our model-fitting procedure simply involved a few iterations of adjusting the parameters by hand until a reasonably good fit was achieved. Because future model fitting will involve a computer program that searches for the exact parameter values that maximize a model's degree of fit and the quality of that fit will depend on the number of free parameters, it is worth considering means by which the number of parameters could be reduced still further. For example, the strength of semantic relationships relating features and prior concepts could be independently measured in the form of subject ratings. Alternatively, one could assume that those connection strengths reflect the empirical regularities in the environment in which the model is assumed to have developed and then independently measure those regularities. For example, connection strengths could be set according to how frequently two concepts co-occur in a text corpus (Landauer & Dumais, 1997). Finally, the model could learn the connection strengths itself by first training it on a large text corpus, an approach adopted by many of the language-processing models mentioned above.

Conclusion

We have presented a model of category learning that uses both empirical experience and prior knowledge to form new categories. The model does a good job in qualitatively reproducing a number of results from studies of how knowledge influences category learning. We have suggested extensions to the model that allow it to incorporate more sophisticated forms of knowledge representation and to account for a wider range of empirical-learning phenomena.

REFERENCES

Ahn, W.-K. (1991). Effects of background knowledge on family resemblance sorting and missing features. In K. J. Hammond & D. Gentner (Eds.), *Proceedings of the Thirteenth Annual Conference of the Cognitive Science Society* (pp. 203-208). Hillsdale, NJ: Erlbaum.

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

Ahn, W.-K., Kim, N. S., Lassaline, M. E., & Dennis, M. J. (2000). Causal status as a determinant of feature centrality. *Cognitive Psychology*, **41**, 361-416.

Almeida, L. B. (1987). A learning rule for asynchronous perceptrons with feedback in a combinatorial environment. In M. Caudil & C. Butler (Eds.), *Proceedings of the IEEE First International Conference on Neural Networks* (pp. 609-618). Los Alamitos, CA: IEEE Computer Society Press.

Anderson, J. A., & Murphy, G. L. (1986). Concepts in connectionist models. In J. S. Denker (Ed.), *Neural networks for computing* (pp. 17-22). New York: American Institute of Physics.

Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., & Waldron, E. M. (1998). A neuropsychological model of multiple systems in category learning. *Psychological Review*, **105**, 442-481.

Ashby, F. G., & Waldron, E. M. (1999). On the nature of implicit categorization. *Psychonomic Bulletin & Review*, **6**, 363-378.

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

Brachman, R. J. (1979). On the epistemological status of semantic networks. In N. V. Findler (Ed.), *Associative networks: Representation and use of knowledge by computers* (pp. 3-50). New York: Academic Press.

Brewer, W. F., & Nakamura, G. V. (1984). The nature and functions of schemas. In S. W. Robert & K. S. Thomas (Eds.), *Handbook of social cognition* (pp. 119-160). Hillsdale, NJ: Erlbaum.

Catrambone, R., & Holyoak, K. J. (1989). Overcoming contextual limitations on problem-solving transfer. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **15**, 1147-1156.

Choi, S., McDaniel, M. A., & Busemeyer, J. R. (1993). Incorporat-

ing prior biases in network models of conceptual rule learning. *Memory & Cognition*, **21**, 413-423.

CHRISTIANSEN, J. H., & CHATER, N. (1999). Toward a connectionist model of recursion in human linguistic performance. *Cognitive Science*, **23**, 157-205.

DELL, G. S., JULIANO, C., & GOVINDJEE, A. (1993). Structure and content in language production: A theory of frame constraints in phonological speech errors. *Cognitive Science*, **17**, 149-195.

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

ESTES, W. K. (1994). *Classification and cognition*. New York: Oxford University Press.

FODOR, J. A., & PYLYSHYN, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. In S. Pinker & J. Mehler (Eds.), *Connections and symbols* (pp. 3-72). Cambridge, MA: MIT Press, Bradford Books.

FRANKS, J. J., & BRANSFORD, J. D. (1971). Abstraction of visual patterns. *Journal of Experimental Psychology*, **90**, 65-74.

GASKELL, M. G., & MARSLEN-WILSON, W. D. (1997). Integrating form and meaning: A distributed model of speech perception. *Language & Cognitive Processes*, **12**, 613-656.

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

GOLDSTONE, R. L. (1994). Influences of categorization on perceptual discrimination. *Journal of Experimental Psychology: General*, **130**, 116-139.

GOLDSTONE, R. L. (1996). Isolated and interrelated concepts. *Memory & Cognition*, **24**, 608-628.

GOLDSTONE, R. L. (2000). Unitization during category learning. *Journal of Experimental Psychology: Human Perception & Performance*, **26**, 86-112.

GOLDSTONE, R. L., & STEYVERS, M. (2001). The sensitization and differentiation of dimensions during category learning. *Journal of Experimental Psychology: General*, **123**, 116-139.

GOLDSTONE, R. L., STEYVERS, M., SPENCER-SMITH, J., & KERSTEN, A. (2000). Interactions between perceptual and conceptual learning. In E. Diettrich & A. B. Markman (Eds.), *Cognitive dynamics: Conceptual and representational change in humans and machines* (pp. 189-228). Mahwah, NJ: Erlbaum.

HAMPTON, J. A. (1979). Polymorphous concepts in semantic memory. *Journal of Verbal Learning & Verbal Behavior*, **18**, 441-461.

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. (1997). Knowledge and concept learning. In K. Lamberts & D. Shanks (Eds.), *Knowledge, concepts, and categories* (pp. 7-42). Cambridge, MA: MIT Press.

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. (2001). Background knowledge and models of categorization. In U. Hahn & M. Ramscar (Eds.), *Similarity and categorization* (pp. 155-178). Oxford: Oxford University Press.

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.

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

HINTON, G. E., & MCCLELLAND, J. L. (1988). Learning representations by recirculation. In D. Z. Anderson (Ed.), *Neural information processing systems* (pp. 358-366). New York: American Institute of Physics.

HINTON, G. E., & SEJNOWSKI, T. J. (1986). Learning and relearning in Boltzmann machines. In D. E. Rumelhart, J. L. McClelland, & the PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (pp. 282-317). Cambridge, MA: MIT Press.

HINTON, G. E., & SHALLICE, T. (1991). Lesioning an attractor network: Investigations of acquired dyslexia. *Psychological Review*, **98**, 74-95.

HOLYOAK, K. J. (1991). Symbolic connectionism: Toward third-generation theories. In K. A. Ericsson & J. Smith (Eds.), *Toward a general theory of expertise: Prospects and limits* (pp. 301-336). Cambridge: Cambridge University Press.

HOLYOAK, K. J., & SIMON, D. (1999). Bidirectional reasoning in decision making by constraint satisfaction. *Journal of Experimental Psychology: General*, **128**, 3-31.

HOLYOAK, K. J., & THAGARD, P. (1989). Analogical mapping by constraint satisfaction. *Cognitive Science*, **13**, 295-355.

HOPFIELD, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, **81**, 3088-3092.

HULL, C. L. (1920). Quantitative aspects of the evolution of concepts. *Psychological Monographs*, **28**, 1-86.

HUMMEL, J. E., & HOLYOAK, K. J. (1997). Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review*, **104**, 427-466.

KAMIN, L. J. (1969). Predictability, surprise, attention, and conditioning. In R. Church & B. A. Campbell (Eds.), *Punishment and aversive behavior* (pp. 261-279). New York: Appleton-Century-Crofts.

KAPLAN, A. S., & MURPHY, G. L. (1999). The acquisition of category structure in unsupervised learning. *Memory & Cognition*, **27**, 699-712.

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

KEIL, F. C. (1995). The growth of causal understandings of natural kinds. In D. Sperber, D. Premack, & A. J. Premack (Eds.), *Causal cognition: A multidisciplinary approach* (pp. 234-262). Oxford: Oxford University Press, Clarendon Press.

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

KRUSCHKE, J. K. (1996a). Base rates in category learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **27**, 3-26.

KRUSCHKE, J. K. (1996b). Dimensional relevance shifts in category learning. *Connection Science*, **8**, 201-223.

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

KRUSCHKE, J. K., & BLAIR, N. J. (2000). Blocking and backward blocking involve learned inattention. *Psychonomic Bulletin & Review*, **7**, 636-645.

KRUSCHKE, J. K., & JOHANSEN, M. K. (1999). A model of probabilistic category learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **25**, 1083-1119.

KUNDA, Z., & THAGARD, P. (1996). Forming impressions from stereotypes, traits, and behaviors: A parallel-constraint satisfaction theory. *Psychological Review*, **103**, 284-308.

LANDAUER, T. K., & DUMAIS, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of knowledge acquisition, induction, and representation. *Psychological Review*, **104**, 211-240.

LIN, E. L., & MURPHY, G. L. (1997). The effects of background knowledge on object categorization and part detection. *Journal of Experimental Psychology: Human Perception & Performance*, **23**, 1153-1163.

MARCUS, G. F. (2001). *The algebraic mind: Integrating connectionism and cognitive science*. Cambridge, MA: MIT Press.

MCCLELLAND, J. L., & RUMELHART, D. E. (1981). An interactive activation model of context effects in letter perception: Pt. 1. An account of basic findings. *Psychological Review*, **88**, 375-407.

MCCLELLAND, J. L., & RUMELHART, D. E. (1985). Distributed memory and the representation of general and specific information. *Journal of Experimental Psychology: General*, **114**, 159-188.

MCLEOD, P., SHALLICE, T., & PLAUT, D. C. (2000). Attractor dynamics in word recognition: Converging evidence from errors by normal subjects, dyslexic patients and a connectionist model. *Cognition*, **74**, 91-113.

MEDIN, D. L., & SCHAFFER, M. M. (1978). Context theory of classification learning. *Psychological Review*, **85**, 207-238.

MEDIN, D. L., & SCHWANENFLUGEL, P. J. (1981). Linear separability in

classification learning. *Journal of Experimental Psychology: Human Learning & Memory*, **7**, 355-368.
MEDIN, D. L., & SHOBEN, E. J. (1988). Context and structure in conceptual combination. *Cognitive Psychology*, **20**, 158-190.
MOVELLAN, J. R. (1989). Contrastive Hebbian learning in the continuous Hopfield model. In D. S. Touretzky, G. E. Hinton, & T. J. Sejnowski (Eds.), *Proceedings of the 1988 Connectionist Models Summer School* (pp. xxx-xxx). San Mateo, CA: Kaufmann.
MURPHY, G. L. (1993). Theories and concept formation. In I. Van Mechelen, J. Hampton, R. S. Michalski, & P. Theuns (Eds.), *Categories and concepts: Theoretical views and inductive data analysis* (pp. 173-200). London: Academic Press.
MURPHY, G. L. (2002). *The big book of concepts*. 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.
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. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **10**, 104-114.
NOSOFSKY, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology*, **115**, 39-57.
NOSOFSKY, R. M., & PALMERI, T. J. (1997). An exemplar-based random walk model of speeded classification. *Psychological Review*, **104**, 266-300.
NOSOFSKY, R. M., PALMERI, T. J., & MCKINLEY, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review*, **101**, 53-79.
O´REILLY, R. C. (1996). Biologically plausible error-driven learning using local activation differences: The generalized recirculation algorithm. *Neural Computation*, **8**, 895-938.
PALMERI, T. J., & BLALOCK, C. (2000). The role of background knowledge in speeded perceptual categorization. *Cognition*, **77**, B45-B47.
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.
PEARLMUTTER, B. A. (1995). Gradient calculations for dynamic recurrent networks: A survey. *IEEE Transactions on Neural Networks*, **6**, 1212-1228.
PINEDA, F. J. (1987). Generalization of backpropagation to recurrent neural networks. *Physical Review Letters*, **19**, 2229-2232.
PLAUT, D. C., & BOOTH, J. R. (2000). Individual and developmental differences in semantic priming: Empirical and computational support for a single-mechanism account of lexical processing. *Psychological Review*, **107**, 786-823.
POSNER, M. I., & KEELE, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, **77**, 353-363.
PROFFITT, J. B., COLEY, J. D., & MEDIN, D. L. (2000). Expertise and category-based induction. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **26**, 811-828.
READ, S. J., & MILLER, L. C. (1994). Dissonance and balance in belief systems: The promise of parallel constraint satisfaction processes and connectionist modeling approaches. In R. C. Schank & E. J. Langer (Eds.), *Belief, reasoning, and decision making: Psycho-logic in honor of Bob Abelson* (pp. 209-235). Hillsdale, NJ: Erlbaum.
REHDER, B. (2003a). Categorization as causal reasoning. *Cognitive Science*, **27**, 709-748.
REHDER, B. (2003b). A causal-model theory of conceptual representation and categorization. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **29**, 1141-1159.
REHDER, B., & HASTIE, R. (2001). Causal knowledge and categories: The effects of causal beliefs on categorization, induction, and similarity. *Journal of Experimental Psychology: General*, **130**, 323-360.
REHDER, B., & HASTIE, R. (in press). Category coherence and category-based property induction. *Cognition*.
ROSCH, E. H., & MERVIS, C. B. (1975). Family resemblance: Studies in the internal structure of categories. *Cognitive Psychology*, **7**, 573-605.
ROSS, B. H., & KENNEDY, P. T. (1990). Generalizing from the use of earlier examples in problem solving. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **16**, 42-55.
ROSS, B. H., & MURPHY, G. L. (1999). Food for thought: Cross-classification and category organization in a complex real-world domain. *Cognitive Psychology*, **38**, 495-553.
RUMELHART, D. E. (1980). Schemata: The building blocks of cognition. In R. J. Spiro, B. C. Bruce, & W. F. Brewer (Eds.), *Theoretical issues in reading comprehension* (pp. 64-80). Hillsdale, NJ: Erlbaum.
RUMELHART, D. E., & MCCLELLAND, J. L. (1982). An interactive activation model of context effects in letter perception: Pt. 2. The contextual enhancement effect and some tests and extensions of the model. *Psychological Review*, **89**, 60-94.
RUMELHART, D. E., & MCCLELLAND, J. L. (1986). *Parallel distributed processing: Exploration in the microstructure of cognition*. Cambridge, MA: MIT Press.
SCHYNS, P. G., GOLDSTONE, R. L., & THIBAUT, J. (1998). The development of features in object concepts. *Behavioral & Brain Sciences*, **21**, 1-54.
SCHYNS, P. [G.], & MURPHY, G. L. (1994). The ontogeny of part representation in object concepts. In D. L. Medin (Ed.), *The psychology of learning and motivation* (Vol. 31, pp. 305-349). San Diego: Academic Press.
SCHYNS, P. G., & RODET, L. (1997). Categorization creates functional features. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **23**, 681-696.
SHASTRI, L., & AJJANAGADDE, V. (1993). From simple associations to systematic reasoning: A connectionist representation of rules, variables, and dynamic bindings using temporal synchrony. *Behavioral & Brain Sciences*, **16**, 417-494.
SHULTZ, T. R., & LEPPER, M. R. (1996). Cognitive dissonance reduction as constraint satisfaction. *Psychological Review*, **103**, 219-240.
SLOMAN, S., LOVE, B. C., & AHN, W.-K. (1998). Feature centrality and conceptual coherence. *Cognitive Science*, **22**, 189-228.
SMITH, E. E., & MEDIN, D. L. (1981). *Categories and concepts*. Cambridge, MA: Harvard University Press.
SMITH, E. E., PATALANO, A. L., & JONIDES, J. (1998). Alternative strategies of categorization. *Cognition*, **65**, 167-196.
SMOLENSKY, P. (1986). Information processing in dynamical systems: Foundations of harmony theory. In D. E. Rumelhart, J. L. McClelland, & the PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (pp. 194-281). Cambridge, MA: MIT Press.
SMOLENSKY, P. (1990). Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence*, **46**, 259-310.
SPALDING, T. L., & MURPHY, G. L. (1996). Effects of background knowledge on category construction. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **22**, 525-538.
SPALDING, T. L., & MURPHY, G. L. (1999). What is learned in knowledge-related categories? Evidence from typicality and feature frequency judgments. *Memory & Cognition*, **27**, 856-867.
SPELLMAN, B. A., & HOLYOAK, B. A. (1993). A coherence model of cognitive consistency: Dynamics of attitude change during the Persian Gulf War. *Journal of Social Issues*, **49**, 147-165.
TABOR, W., CORNELL, J., & TANENHAUS, M. K. (1997). Parsing in a dynamical system: An attractor-based account of the interaction of lexical and structural constraints in sentence processing. *Language & Cognitive Processes*, **12**, 211-271.
THAGARD, P. (1989). Explanatory coherence. *Behavioral & Brain Sciences*, **12**, 435-502.
THAGARD, P., & MILLGRAM, E. (1995). Inference to the best plan: A coherence theory of decision. In A. Ram & D. B. Leake (Eds.), *Goal-driven learning* (pp. 439-454). Cambridge, MA: MIT Press, Bradford Books.
TOURETZKY, D. S., & HINTON, G. E. (1988). A distributed connectionist production system. *Cognitive Science*, **12**, 423-466.
WALDMANN, M. R., HOLYOAK, K. J., & FRATIANNE, A. (1995). Causal

models and the acquisition of category structure. *Journal of Experimental Psychology: General*, **124**, 181-206.

WATTENMAKER, W. D., DEWEY, G. I., MURPHY, T. D., & MEDIN, D. L. (1986). Linear separability and concept learning: Context, relational properties, and concept naturalness. *Cognitive Psychology*, **18**, 158-194.

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

WISNIEWSKI, E. J., & MEDIN, D. L. (1994). On the interaction of theory and data in concept learning. *Cognitive Science*, **18**, 221-282.

ZIPSER, D. (1986). Biologically plausible models of place recognition and goal location. In D. E. Rumelhart, J. L. McClelland, & the PDP Research Group (Eds.), *Parallel distributed processing: Exploration in the microstructure of cognition* (pp. 432-470). Cambridge, MA: MIT Press.

NOTES

1. The sequential updating of units within a cycle only approximates the intended parallel updating of units in a constraint satisfaction network. In order to approximate parallel updating more closely, each unit's activation function was adjusted to respond more slowly to its total input. Specifically, in cycle i, a unit's activation was updated according the function $act_i = 1 / 1+ \exp(-adj\text{-}input_i)$, where $adj\text{-}input_i$ is a weighted average of the adjusted input from the previous cycle and the total input from the current cycle. Specifically,

$$adj\text{-}input_i = adj\text{-}input_{i-1} + (adj\text{-}input_i - adj\text{-}input_{i-1})/gain.$$

In the present simulations, $gain = 4$.

2. Because the output units are sigmoid units, a positive external input to the correct category label moves the activation of that unit closer to 1, whereas a negative external input moves the activation of the incorrect category label closer to 0. During the plus phase, the activation of those units could become arbitrarily close to 1 and 0, respectively, by increasing the magnitude of the external input beyond its current value of 1.

3. Although here we emphasize KRES's strengths as an empirical-learning system, it should be noted that there exists some standard learning effects that it is unable to account for in its current state. For example, in addition to the prototype effect just described, Posner and Keele (1968) found that exemplars that were part of the original training set were classified more accurately than the prototype, a result that supports exemplar theories of classification (Medin & Schaffer, 1978; Nosofsky, 1986) and is not predicted by KRES. In the General Discussion section, we review KRES's successes and failures as an empirical-learning system and discuss extensions that address some of its deficiencies.

4. For consistent terminology across simulations, we use *related* to refer to conditions that have prior knowledge and to the features that are related (via that prior knowledge) to other features or concepts. We use *unrelated* to refer to conditions with no prior knowledge and to the features that are unrelated to other features or concepts. In the original articles reporting these experiments, a variety of terms for those conditions were used.

5. In the mixed-theme condition, half a category's idiosyncratic features were related to one theme, and the other half to another theme. However, Kaplan and Murphy (2000) found that performance in this condition did not differ significantly from a no-theme condition in which there were no themes linking idiosyncratic features (Experiment 3). Hence, we omit any feature–feature relationships in our simulation of the mixed-theme (unrelated) condition reported below.

6. For example, we have implemented a version of KRES with exemplar nodes that are connected to their constituent features and that become active via recurrent connections when those features are active. When the connection weights between the exemplar nodes and the category label units are learned according to CHL, such a network solves XOR problems easily. In addition, KRES can solve XOR and other nonlinearly separable problems with hidden units placed between the feature units and the category nodes (see O'Reilly, 1996, for a demonstration of solving XOR problems using CHL and a recurrent network with hidden units).

7. There is also nothing in the model to prevent other sorts of feature relationships or structures from being incorporated. For example, if representing causal knowledge as an asymmetric relation turns out to be important (as has been proposed by Ahn, 1998, Rehder, 2003a, 2003b, and Sloman et al., 1998), such relations can be incorporated into KRES as (say) two unidirectional links with different weights, one from cause to effect and the other from effect to cause.

8. Note that even for the highly similar exemplars that are retrieved, there is no guarantee that the *predictive relations* among features will be appropriate for the new category. For example, although subjects may well have been reminded of other vehicles when learning about arctic vehicles (because arctic vehicles are more similar to other vehicles than they are to buildings or clothing), many vehicles possess feature combinations that are incorrect for the *arctic vehicle* category (e.g., Volvos do not drive on glaciers, snowmobiles are not insulated, Zambonis are not made in Norway, etc.).

9. Although it is not hard to imagine such mechanisms. In KRES, for example, feature units could activate exemplar units (see note 6), which would then activate the category label unit(s) to which they are associated (by previous learning). CHL would then be used in the usual manner to learn the connections from these prior category units and exemplars to the new category units. Note that this proposed model is similar to the models with prior concepts that we used in Simulations 2 and 3, with the exception that activation of a prior concept is mediated by one or more of its own exemplars. Another possibility is that prior exemplars aid learning because previous category learning has resulted in their being strongly encoded in memory.

(Manuscript received June 19, 2001;
revision accepted for publication August 20, 2002.)