Modeling Category Learning with Exemplars and Prior Knowledge

Harlan D. Harris (harlan.harris@nyu.edu)
Bob Rehder (bob.rehder@nyu.edu)
New York University, Department of Psychology
New York, NY 10003 USA

Abstract

An open question in category learning research is how prior knowledge affects the process of learning new concepts. Rehder and Murphy's (2003) Knowledge Resonance (KRES) model of concept learning uses an interactive neural network to account for many observed effects related to prior knowledge, but cannot account for the learning of nonlinearly separable concepts. In this work, we extend the KRES model by adding exemplar nodes. The new model accounts for the fact that linearly separable concepts are not necessarily easier than nonlinearly separable concepts (Medin & Schwanenflugel, 1981), and more importantly, accounts for a notable interaction between the presence of useful prior knowledge and linear separability (Wattenmaker, Dewey, Murphy, & Medin, 1986). Two architectural variants of the model were tested, and the dependence of good results on a particular architecture, indicates how formal modeling can uncover facts about how the prior knowledge which influences concept learning is used and represented.

Most current theories of category learning address how new concepts are learned on the basis of empirical regularities in the environment. Considerable progress has been made in determining how learners encode empirical information about how features, and sets of features, covary with category labels. However, these models fail to account for the important role of prior knowledge. Other models of category learning address the effects of prior knowledge, but they in turn fail to account for the wide range of empirical effects that have been observed. The work reported here aims to integrate these two veins of concept learning research.

Prior knowledge is known to have a number of effects on concept learning. When knowledge is related to a learning task, learning is often faster (Murphy & Allopenna, 1994; Wattenmaker et al., 1986). In addition, when new concepts are related to prior knowledge, structural effects that have been found in empirical concept learning studies may not be found or may even be reversed (Pazzani, 1991; Wattenmaker et al., 1986).

In this research we introduce a new category learning model whose goal is to account for effects of both prior knowledge and empirical regularities on concept learning. We address the question of which kinds of representations (exemplars? prototypes? rules?) are involved in learning tasks and how those representations become related to one another and to representations of prior knowledge as a result of experience. By fitting variants of our new model to two human learning data sets, we will show that only a very particular pattern of connectivity among representations is warranted.

We pursue this question by extending an existing model of category learning, the Knowledge Resonance (KRES) model introduced by Rehder and Murphy (2003). KRES is a connectionist model of knowledge effects in concept learning that uses interactive activation among representations of stimulus features, category labels, and prior knowledge, then uses a supervised learning algorithm called Contrastive Hebbian Learning (CHL: O'Reilly, 1996) to learn symmetrical weights between the representations. KRES accounts for effects of prior knowledge on learning rate, generalization patterns and reaction time.

KRES builds on the Baywatch model introduced by Heit and Bott (2000). Baywatch is a standard feedforward connectionist network supplemented with prior concept nodes that can be used as the basis for categorization. Baywatch accounts for knowledge effects on responses to novel but knowledge-related features, as well as to prior knowledge that is incongruent with the empirical stimuli (Heit, Briggs, & Bott, 2004). KRES goes beyond Baywatch in being able to also represent prior knowledge that relates stimulus features to (a) one another, and (b) concept nodes (allowing the model to account for "top down" effects in learning).

Nevertheless, the published versions of both Baywatch and KRES have a significant restriction. They are limited to learning linearly separable concepts. Their architectures are similar to a classical prototype model, where the weights compute a monotonic function of the input representation. However, people are able to learn nonlinearly separable concepts, and often find such concepts as easy to learn as linearly separable ones (Medin & Schwanenflugel, 1981). In part as a result of this, exemplar models of classification (Medin & Schaffer, 1978; Nosofsky, 1986; Kruschke, 1992) have become prominent, as they naturally account for learning of nonlinearly separable concepts. Some recent work has challenged exemplar models (Smith & Minda, 1998, but see Nosofsky & Zaki, 2002; Rehder & Hoffman, 2005), and some recent models of classification have proposed alternatives, such as using clusters instead of exhaustive sets of exemplars (Love, Medin, & Gureckis, 2004). However, our goal is to build a new model with the ability to learn nonlinearly separable concepts, and exemplars are a reasonable starting place with much empirical evidence to

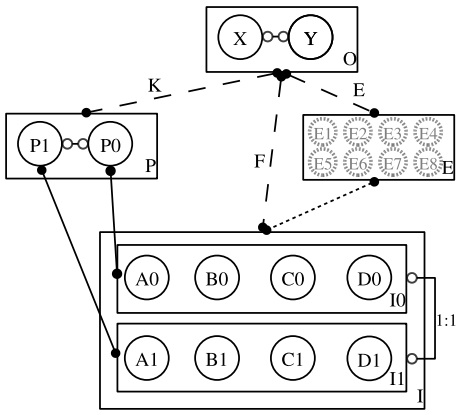

Figure 1: Architecture of new KRES network. I = input nodes; O = output nodes; P = prior knowledge nodes; E = exemplar nodes. Connections depicted with solid lines are fixed weights; with fine dashed lines are set-once exemplar weights; with dashed lines are CHL-learnable weights. The KRES/EFK model includes all three of the links labeled E, F, and K; other models include subsets.

support them.

We first describe how we incorporated exemplars into KRES, noting several possible architectural variations. We then give the results of our work simulating experiments by Medin and Schwanenflugel (1981) and Wattenmaker et al. (1986), focusing on the architectural variations and their implications for theories of concept learning. We conclude with a discussion of the results and their implications for a comprehensive theory of category learning with and without prior knowledge.

New Models

Our new models are simple extensions of the KRES model (Rehder & Murphy, 2003). Figure 1 shows the overall architecture. The models work as follows.

In KRES, connections between nodes are bidirectional, and those connections can either be fixed inhibitory, fixed excitatory, or learned with experience. In Figure 1, there are fixed inhibitory connections between the two prior-knowledge nodes in layer P, fixed excitatory connections between each of the prior-knowledge nodes and one of the two banks of input nodes (e.g., between P0 and I0), learned connections between the two prior-knowledge nodes and the output nodes in layer O, etc. Note that the input (layer I) is represented as pairs of mutually exclusive values of a particular attribute, so nodes A0 and A1 also have fixed inhibitory connections. To make a categorization prediction, the inputs to the model have constant signals applied to them. Activation then spreads throughout the model, both forward and backwards. For example, if the I0 input nodes are active, that activity will resonate with the prior knowledge P0 node, and their activation will increase as a result. Activation of nodes is a sigmoidal function of the total input, with a steepness parameter α. After many cycles of spreading activation, the network settles, and the activations of the output nodes are transformed using a Luce choice rule into the probability that the input is categorized as an X or a Y. The model learns when a teaching signal is then applied to the output nodes. For details, see Rehder and Murphy (2003).

To account for the learning of nonlinearly separable concepts, we added exemplar nodes to KRES. As with ALCOVE (Kruschke, 1992) and related models, we used fixed exemplar nodes activated by matching stimuli. The first time a new exemplar is seen, an exemplar node is connected to the input nodes with fixed weights set to match the stimuli and scaled by a parameter, w_e. The new node is also connected to the output nodes with weights initialized and learned in the usual manner for KRES. Activation of the new exemplar nodes then proceeds as with any other node in a KRES network.

As shown in Figure 1, there are a number of possible variations of the new network's connectivity. In the original KRES model, there were connections between the feature nodes (layer I) and the output nodes (layer O), and the prior-knowledge nodes (layer P if present) and the output nodes. With the addition of the new exemplar nodes (layer E), the full model, which we notate as KRES/EFK, has three separate possible bases for classification: the exemplar, feature, and knowledge nodes. However, each of those connections is theoretically optional (although one must be present), and each is theoretically interesting. The exemplar connections are essential if nonlinearly separable concepts are to be learned, so we did not consider variations without them.

The KRES/E model has only these exemplar connections, and no other connections to the output nodes. The feature nodes influence categorization only by activating the exemplar nodes, while the prior-knowledge nodes (if present) can influence categorization only by influencing the activation of the input nodes. Note that in KRES/E the sole effect of the prior concept nodes is to modify the activation of the input features so that they are more consistent with those concepts. For example, a stimulus with many P0-linked features will strongly activate P0, which in turn will reinforce those features (and thus dampen the activation of any features associated with P1). The feature nodes, thus partially canonicalized (i.e., made more consistent with P0), will then exert their influence on the category labels via the exemplar nodes.

In the KRES/EK model, output nodes have connections from both the exemplar and the prior knowledge nodes. The prior knowledge concepts not only modulate input feature activation, they can also be used as the basis for categorization, with (potentially) no contribution from the exemplar nodes.

The KRES/EF model has connections from the exemplar nodes as well as direct connections from the input features. It can be seen as a model that combines exemplar and prototype-like computations in its effort to categorize. As with KRES/E, prior-knowledge nodes can influence the activation of the input nodes, but cannot be directly used as the basis for categorization.

Table 1: Category structure for Preliminary Simulation, based on Medin & Schwanenflugel (1981, Experiment 1).

Linearly Separable		Non-linearly Separable	
Stimuli	Category	Stimuli	Category
1011	A	1000	A
1010	A	0111	A
1101	A	1110	A
0110	A	1011	A
1001	B	0110	B
0010	B	1001	B
0100	B	0000	B
0001	B	0001	B

KRES/EFK has all three sets of connections to the output nodes, allowing categorization decisions to be made with any combination of prior concept, input, or exemplar activations.

By comparing these models, we hope to show that a particular pattern of connectivity among representations is needed to account for the experimental findings.

Preliminary Simulation: Nonlinearly separable concepts

We begin by confirming that our new model can indeed learn nonlinearly separable concepts, and (more ambitiously) by testing whether one of its variants exhibits the same learning patterns as people. In their Experiment 1, Medin and Schwanenflugel (1981) notably found that linearly separable categories were not necessarily easier to learn than nonlinearly separable categories. This highly influential result was one of several that undermined the independent-cue, or prototype, models of category learning. We investigated whether one of the variants of our new model would reproduce the equivalent learning difficulty of Medin and Schwanenflugel's linearly separable and nonseparable category structures.

The new model, with exemplar nodes but without prior knowledge, was trained on the two conditions shown in Table 1. Both KRES/E and KRES/EF (see Figure 1) were fit. (Because of the absence of prior knowledge, KRES/EK reduces to KRES/E for this task.) Three parameters were explored systematically. These were the learning rate, LR, the strength of the fixed inhibitory weights, w_{in}, and α, the sharpness of the sigmoidal squashing function. (High values of α force node activations to be either very high or very low.) The strength of exemplar weights, w_e, was fixed at 1. The bias on the exemplar nodes, b_e, was set to be a function of α such that the activation of the exemplar nodes was $\frac{1}{n}$ (where n is the number of exemplars) when the net input to a node was 0: $b_e = -\frac{\log(n-1)}{\alpha}$. Other parameters, such as the gain, remained at the defaults reported in Rehder and Murphy (2003).

A parameter search was performed, with replications to reduce effects of noise. Each replication involved a run of the model, learning the stimuli shown in Table 1, with the same number of blocks as in the experimental work. Erroneous predictions during training were counted for each item. The results were compared with the per-item error rates reported by Medin and Schwanenflugel (1981). Overall, the strength of inhibitory weights was not critical, as long as they were adequately inhibitory. We thus set the strength of inhibitory weights to $w_{in} = -2$ for the simulations reported here.

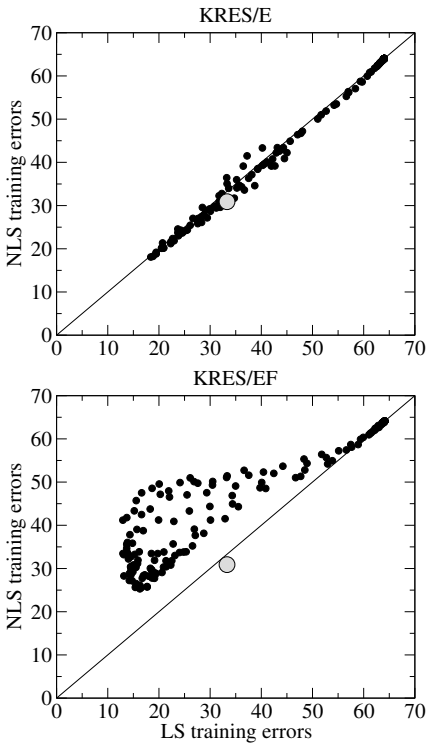

Figure 2: Performance of two KRES variants on Medin & Schwanenflugel (1981) task, showing overall error counts with various parameter settings. Grey circles are experimental error counts. Chance performance is 64 errors.

For KRES/EF, the best results were found with $LR = 0.1, \alpha = 1.5$. The MSE and χ^2 error, relative to the experimental per-item error rates, were 2.84 and 11.41, respectively. For KRES/E, the best results were found with $LR = 0.4, \alpha = 1.2$. The MSE and χ^2 error were much lower, at 0.66 and 2.59.

Medin and Schwanenflugel (1981) reported a mean 33.3 errors on the LS problems and 30.9 errors on the NLS problems. KRES/E made about the same number of errors on the LS and NLS problems (32.6 and 31.6, respectively), while KRES/EF showed easier learning of the linearly separable category (26.6 versus 35.2 errors). Figure 2 shows a scatter plot of the number of training errors on the two problems for the two models, over a wide range of parameter settings. KRES/E accounts for the qualitative Medin and Schwanenflugel (1981) result, regardless of parameter settings; for KRES/E, the two concepts are about equally difficult. For KRES/EF, however, no parameter settings were able to reproduce this result; for KRES/EF, the LS problem is easier, regardless of parameter settings.

Table 2: Category structure for Main Simulation, based on Watternmaker et al. (1986).

Linearly Separable		Non-linearly Separable	
Stimuli	Category	Stimuli	Category
1110	A	1000	A
1011	A	1010	A
1101	A	1111	A
0111	A	0111	A
1100	B	0001	B
0001	B	0100	B
0110	B	1011	B
1010	B	0000	B

KRES/E is successful because exemplar nodes allow equally rapid learning of linearly and nonlinearly separable concepts, and direct connections from input to output nodes are not present. Other exemplar models of classification, such as ALCOVE, share this architecture. (We have also successfully fit ALCOVE to the Medin & Schwanenflugel, 1981 results.) The next simulation, however, goes beyond ALCOVE's scope as a model of concept learning.

Main Simulation: Prior knowledge and linear separability

Wattenmaker et al. (1986) showed an interaction between category structure and prior knowledge. Subjects learned either the linearly separable or nonlinearly separable structure shown in Table 2. In the knowledge-related condition but not the knowledge-unrelated (control) condition the features were correlated with personality traits (e.g., the "1" and "0" features were behaviors which exemplified "honesty" and "dishonesty," respectively). In the knowledge-unrelated condition, Wattenmaker et al. found that the two category structures were about equally easy to learn (there were nonsignificantly fewer errors for the nonlinearly separable structure, Figure 3). When knowledge was present, both structures became easier to learn, illustrating that prior knowledge can speed learning. Importantly however, this effect was stronger with the linearly-separable categories (Figure 3). Apparently, the prior knowledge that the category features independently instantiated known personality traits biased learners toward a "summing" strategy consistent with a linearly-separable concept but less helpful for a nonlinearly-seperable concept (also see Murphy & Kaplan, 2000).

KRES was able to account for the speedup due to prior knowledge in the linearly separable case (Rehder & Murphy, 2003, Simulation 2), but it could not even attempt to account for the interaction with the nonlinearly separable case. The present simulation used both prior concept nodes and exemplar nodes to account for this interaction.

We focused on three of the six possible variants on KRES: KRES/E, KRES/EK, and KRES/EFK. (As shown above, KRES/EF was unable to account for the basic pattern of results in the Medin and Schwanenflugel (1981) data, and thus was not considered.) That is, in addition to accounting for the Wattenmaker et al. (1986) results, the Main Simulation examined whether adding direct connections between the output nodes and the prior knowledge and/or feature nodes is necessary.

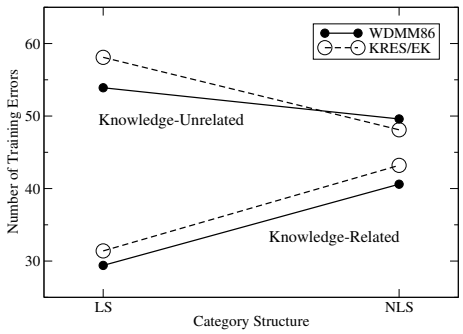

Figure 3: Performance of KRES/EK on Wattenmaker et al. (1986, Experiments 1 and 2) task, compared with experimental results. Chance performance was 64 errors.

As before, the parameter space of the three variations was systematically explored. The learning rate, LR, exemplar weight, w_e, and α were varied, while the other parameters were held constant. The excitatory weights were set to $w_{ex} = 1$ and the inhibitory to $w_{in} = -2$. Training replicated the experimental procedure and replications were performed to get stable estimates of average performance. We combined the results of Wattenmaker et al.'s Experiments 1 and 2 (which had identical category structures but different stimuli), weighted by numbers of subjects, to get a less noisy set of numbers for comparison.

Table 3 shows the best fit for each model, along with the mean squared error. The KRES/EK model was able to closely fit the quantitative and qualitative patterns in the error counts (see Figure 3), while the other models could not. The failures of KRES/EFK to account for these results confirm the model selection results in the first experiment, showing that direct connections from features to the output are not helpful for reproducing the human learning pattern. In addition, the success of KRES/EK relative to KRES/E indicates that connections from prior knowledge nodes to the output nodes *are* helpful. It could have been that prior concepts affected learning in the model merely by changing the activations of the nodes in the input layer, increasing activation levels and pulling the representation towards the prior knowledge prototype. The results of the simulations, however, suggest that this is not the case. The model is unable to account for the experimental results without connections from the prior knowledge nodes to the output nodes.

Per-item fits were quite good for KRES/EK, with a few exceptional points–see Figure 4. We suspect that these exceptions may be partly due to KRES's lack of feature attention weights. The dimensions in Table 2 are not equally diagnostic, and KRES does not shift attention away from less diagnostic dimensions to more diagnostic dimensions.

Table 3: Best fits for the Main Simulation. Parameters, number of training errors, and Mean Squared Error (vs. experimental data) are shown. "Rel" and "Unrel" specify the related (theme) and unrelated (control) conditions.

	LR	w_e	α	Rel / LS	Rel / NLS	Unrel / LS	Unrel / NLS	MSE
WDMM86				29.4	40.6	53.9	49.6	
KRES/E	0.7	1.0	1.0	45.0	32.6	52.3	39.2	105.4
KRES/EK	0.55	0.6	0.9	31.4	43.2	58.1	48.1	7.6
KRES/EFK	0.15	0.6	0.9	36.0	46.6	49.3	49.3	25.3

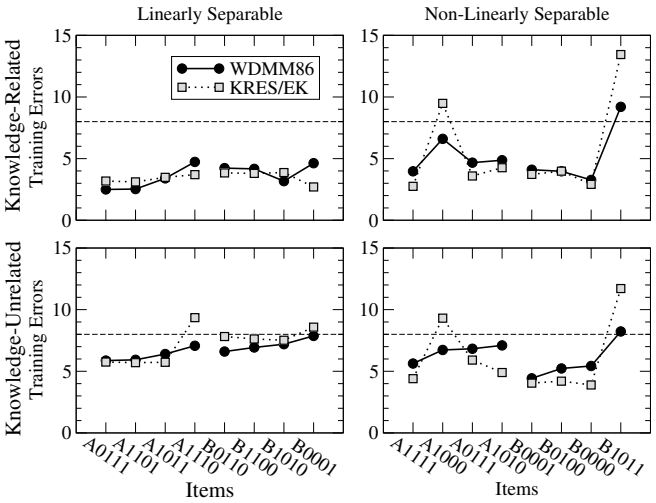

Figure 4: Performance of KRES/EK on Wattenmaker et al. (1986, Experiments 1 and 2) task, showing mean per-item training error rates. The dotted line is chance performance.

The results of the Main Simulation validate the KRES/EK model, showing that it (and not its cousins) can account for the interaction between the presence or absence of prior knowledge and the linear separability of the concept to be learned. Without knowledge, the model finds Wattenmaker et al.'s linearly separable structure slightly more difficult to learn than the nonlinearly separable structure, consistent with the empirical data, and due to the use of exemplar representations. With knowledge, the effect of linear separability reverses, as the model can use prior knowledge nodes directly as the basis for effective learning. Overall, collapsing across category structures, the model accounts for faster learning with prior knowledge, as prior knowledge nodes both directly and indirectly (through influencing representations of feature nodes) aide classification.

Discussion

In this paper we have introduced a new model of concept learning with the potential to account for effects of both empirical regularities and prior knowledge on concept learning. We first showed that KRES/E was able to account for a critical result regarding how empirical regularities affect learning difficulty–nonlinearly categories are not intrinsically more difficult than linearly separable ones (Medin & Schwanenflugel, 1981). Of course, nonlinearly separable category learning is not the only important empirical learning result, but we are confident, based on prior work (Rehder & Murphy, 2003), that the KRES framework exhibits a number of the other standard effects, such as sensitivity to features' category and cue validity and prototype effects. Our new model has thus shown itself to be an empirical learning system faithful to many facets of human learning. We therefore conclude that it is suitable as a platform to model the additional effects of prior knowledge.

We next investigated whether KRES/EK was able to account for the intriguing interaction in the difficulty of learning linear and nonlinear concepts with and without prior knowledge (Wattenmaker et al., 1986). Without knowledge, people found those nonlinearly concepts easier to learn, a preference which was reversed when knowledge was present. Our Main Simulation showed that KRES/EK was able to reproduce this interaction, and even more detailed error results as well. KRES/EK is the only model of category learning able to fully account for these data[1].

We have also shown that successfully accounting for the experimental results involves creating connections among certain kinds of representations and not others. This result is exemplified in our rejection of KRES/EF (in the Preliminary Simulation) and KRES/EFK (in the Main Simulation). When concepts could be built on a basis of raw features, in addition to exemplars and concepts due to prior knowledge, the model was unable to fit the experimental data well. We suggest that this is support for a theory of concept learning where the effect of a stimulus on a categorization decision is mediated by exemplars and prior knowledge. However, the categories studied here both had weak family resemblance structures, which would necessarily impair the usefulness of direct feature-category weights. Future work with other category structures will be needed to confirm that our conclusion holds up more generally.

KRES/EK shares some central architectural assumptions of ALCOVE (Kruschke, 1992). Like ALCOVE, exemplar nodes are used to form the basis for concept learning, with category node activation being a monotonic function of exemplar activation. Of course, KRES/EK extends significantly beyond ALCOVE's scope by being able to incorporate prior knowledge

[1]Heit (2000) uses the integration model of category learning (Heit, 1994), a generalization of Medin and Schaffer's (1978) context model, to account for other results from Wattenmaker et al. (1986).

through prior concept nodes and to account for experimental results with prior knowledge. As discussed in the introduction, KRES/EK also has similarities to Baywatch, with prior knowledge nodes that can act as the basis for categorization (Heit & Bott, 2000). However, our results suggest that Baywatch, without exemplar representations, cannot account for learning of nonlinearly separable concepts, and also cannot account for the Wattenmaker et al. (1986) results. Baywatch is most similar to the versions of KRES with feature-category links, which did not fit the data.

Overall, this work supports a particular view of concept learning and prior knowledge. Categorization is based on a process of parallel constraint satisfaction, where stimuli, prior knowledge and concepts all interact to find the most consistent category response. Prior knowledge can either be based on relationships among features, or (as in this work) can be prior concepts that interact with the stimulus representations and also act as a potential basis for categorization. Exemplars, or perhaps more abstract representations, form the basis for empirical categorization.

Planned future work includes further investigation of the architectural constraints discovered here, of other interactions with knowledge, and of KRES's limitations. Most current models of concept learning include a mechanism for selective attention, and some also can account for unsupervised learning effects, but KRES currently does neither. We also plan to investigate the effects of redundant, irrelevant, and incongruent knowledge. The work described here is a significant step towards developing a truly comprehensive model of concept learning and how it interacts with other aspects of cognition.

Acknowledgements

Support for this research was provided by NIMH Grant MH041704 to Gregory L. Murphy. Thanks to Gregory Murphy and the anonymous reviewers for helpful comments.

References

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

Heit, E. (2000). Background knowledge and models of categorization. In U. Hahn & M. Ramscar (Eds.), *Similarity and categorization.* Oxford University Press.

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

Heit, E., Briggs, J., & Bott, L. (2004). Modeling the effects of prior knowledge on learning incongruent features of category members. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*, 1065–1081.

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

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

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 and Memory*, *7*, 355–368.

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. *The Quarterly Journal of Experimental Psychology*, *53A*, 962–982.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology*, *115*, 39–57.

Nosofsky, R. M., & Zaki, S. R. (2002). Exemplar and prototype models revisited: Response strategies, selective attention, and stimulus generalization. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *28*, 924–940.

O'Reilly, R. C. (1996). Biologically plausible error-driven learning using local activation differences: The generalized recirculation algorithm. *Neural Computation*, *8*, 895–938.

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

Rehder, B., & Hoffman, A. B. (2005). Thirty-something categorization results explained: Attention, eye-tracking, and models of category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 811–829.

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

Smith, J. D., & Minda, J. P. (1998). Prototypes in the mist: The early epochs of category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*, 1411–1436.

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.