An Attention-Based Model of Learning a Function and a Category in Parallel

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

John Paul Minda (jpminda@uwo.ca)
University of Western Ontario, Department of Psychology
London, ON N6A-5C2 Canada

Abstract

Minda and Ross (2004) described two experiments where subjects simultaneously learned both a category and a function. They showed that when both tasks were performed in parallel on the same stimuli, the inductive bias on the categorization task–to focus on a single attribute–spread to the function learning task. Here, we present a new computational model of this phenomenon, using the ALCOVE model of categorization, a new model of function learning, and a hypothesis for their interaction: *shared selective attention.* The model parsimoniously succeeds in learning the category and function, then in accounting for human generalization patterns on conflicting transfer stimuli. The novel function-learning component of the model, extending previous work in mixture-of-experts approaches (Kalish, Lewandowsky, & Kruschke, 2004; Harris & Minda, 2005), is also introduced.

Introduction

Concept learning comprises more than just the classic sort of laboratory supervised training task. Much of the most compelling recent research in concept learning involves other sorts of tasks, such as learning with prior knowledge (Pazzani, 1991), learning through feature induction (Osherson, Smith, Wilkie, López, & Shafir, 1990), learning how to learn effectively (Smith, Jones, Landau, Gershkoff-Stowe, & Samuelson, 2002), learning concepts other than categories (DeLosh, Busemeyer, & McDaniel, 1997), and learning more than one thing at once (Ross, 1997; Minda & Ross, 2004). Here we examine and then model a task that relates to several of these issues.

We define *parallel concept learning* as a situation where a learner has to learn two or more concepts from the same examples. For example Ross (1997, Experiment 1) taught subjects to diagnose (categorize) a disease based on sets of symptoms and in parallel to indicate the appropriate treatment (subcategory). The two tasks were interleaved–for each stimulus presentation, the categorization task and feedback were immediately followed by subcategorization and feedback. Generalization patterns then showed that the concepts learned by the subjects influenced each other.

Minda and Ross (2004) described two parallel concept learning experiments where subjects simultaneously learned both a category, as in a standard concept learning task (Murphy, 2002), and a function, as in a function learning[1] task (DeLosh et al., 1997). They were interested in the inductive biases of these two tasks, and how they interact. As further described below, both concepts could be learned in either of two ways, either by use of a single *criterial attribute* or based on a more general *family resemblance* structure. There is a strong inductive bias to use a unidimensional rule in categorization tasks when possible (Ashby & Ell, 2001; Nosofsky, Palmeri, & McKinley, 1994), but the equivalent inductive bias on the function learning task was unclear. Also in question was whether this bias would be affected by learning the two tasks simultaneously.

In Minda and Ross's (2004) experiments, participants learned a function related to some of the stimuli's features while (for half the subjects) simultaneously learning to categorize the stimuli. For the first task, participants were shown a series of imaginary animals of various sizes and were trained to predict how much food they thought each animal should receive. Correct responses were a function of both size and (implicitly) category membership, where category membership could be determined in one of two different ways by a set of binary features of the animals. (See Table 1.) One way of determining category membership was to use only a single feature while ignoring the rest. Minda and Ross called this a criterial attribute (CA) approach, as that feature was sufficient for correct predictions. The other way of determining category membership was to note the overall similarity structure of the features. This approach was called a family resemblance (FR) strategy. Both strategies can be equally effective, although prior research has shown that FR structures can be difficult to learn in some circumstances. Note that subjects were not told of any underlying structures.

Half of the participants only performed the prediction task, receiving feedback in the form of the correct amount of food for each animal, but receiving no additional information about category membership. The other half of the participants performed this task while simultaneously performing a categorization task. For

[1] As *category learning* is the process of learning how to map from the features of a stimulus to one of a limited number of responses, *function learning* is the analogous process of learning how to map from the features of a stimulus to an unbounded, scalar response. Also, stimuli in function learning usually involves dimensions with ordinal values, while category learning studies tend to use dimensions with nominal values (but see, e.g. Nosofsky & Palmeri, 1996).

Table 1: Training category structure for simulation of Minda and Ross (2004) experiments. The first cue also served as the category label in the Categorization + Prediction condition. The transfer task category structure was the same, except with the first cue reversed.

Category A			Category B		
Cues	Size	Food	Cues	Size	Food
00000	0	4	**1**1111	0	8
01000	0	4	**1**0111	0	8
00100	0	4	**1**1011	0	8
00010	0	4	**1**1101	0	8
00001	0	4	**1**1110	0	8
00000	1	7	**1**1111	1	11
01000	1	7	**1**0111	1	11
00100	1	7	**1**1011	1	11
00010	1	7	**1**1101	1	11
00001	1	7	**1**1110	1	11
00000	2	10	**1**1111	2	14
01000	2	10	**1**0111	2	14
00100	2	10	**1**1011	2	14
00010	2	10	**1**1101	2	14
00001	2	10	**1**1110	2	14

each stimulus, the subjects were first asked to categorize the animal into one of two categories. The categories were equivalent to the values of the criterial attributes, and were also the same as the (implicit) categories learned as part of the first task. After receiving feedback on their categorization, the subjects were then asked to perform the prediction task. The experimental question was whether this interleaving of the prediction task with a related categorization task would change the nature of the concepts learned by the subjects.

After learning, both groups were tested on novel generalization items that were constructed to reveal whether subjects had acquired a concept with the CA or with the FR strategy. The first cue (the CA) was reversed in these transfer stimuli, so that generalization patterns would indicate whether the subject was relying more on that cue or on the FR structure. If the concept was based on the FR structures, then their predictions would be the same as the values shown in Table 1. If the concept was based instead on the CA, and the FR features were ignored, then the predictions would be reversed (8 instead of 4, etc.) See Minda and Ross (2004) for further details. Also, subjects were surveyed to get subjective estimates of how much attention they paid to each stimulus dimension when performing the prediction task.

The results of two versions of this experiment suggested that some participants who learned in the prediction-only condition (performing only function learning, with no categorization) displayed a broader distribution of attention (as measured by generalization patterns as well as the survey results) than participants who learned in parallel to classify the items as well as to make predictions. Participants in the prediction-only group were more likely to use the family resemblance structure of the categories, even with a perfect criterial attribute available. In contrast, participants who also learned to explicitly classify the objects overwhelmingly preferred to learn the criterial attribute (single feature) rules, and few generalization responses showed much evidence of relying on the family resemblance structure.

Minda and Ross's (2004) results suggest that in function learning, the bias to pay attention to a single cue may be weaker than the bias in categorization. When only the function learning (prediction) task was performed, a significant number of subjects spread their attention widely, extracted the FR structure of the cues, and generalized accordingly. When both the function learning and category learning tasks were performed, however, almost all subjects narrowed their attention, focused on the CA in the cues, and generalized accordingly. Thus, although the results are consistent with prior research that has argued that rules are the default approach to category learning (Ashby & Ell, 2001; Nosofsky et al., 1994), that approach does not seem to extend to function learning. The mechanisms underlying these effects are not at all clear. To further understand this task and Minda and Ross's results, we elected to investigate using simulations.

Our hypothesis is that the mechanism for the interaction between the two tasks is simply *shared selective attention.* In a parallel concept learning task, where two learning tasks are performed in parallel using the same stimuli, selective attention is able to focus attention on certain features of the stimuli for both learning tasks. Any tendency one task has to bias attention to particular sets of features will then affect performance on the other task.

In the Components section of this paper we overview the models used in the simulations and how they can be combined in a parallel concept learning task. In the Simulations section we describe the simulations and present the results. We conclude with a Discussion section which contains conclusions and some thoughts on future work.

Components

To model category learning, we chose the well-studied ALCOVE model (Kruschke, 1992). ALCOVE is a neural network model of concept learning, based on exemplar models of categorization (Nosofsky, 1992). It learns to classify stimuli based on monotonic functions of exemplar activations, with error-driven changes to weights and attention. Exemplars are activated based on their similarity to attention-weighted stimulus representations. Our requirements for a category learning model were that the model be clearly described, have selective attention that weights stimulus representations, and that it be a process model of learning. ALCOVE fits our requirements, although some of its particulars (such as its use of a comprehensive set of exemplars as a basis for categorization) are not essential to our simulations, and other alternatives (e.g. clusters as in SUSTAIN, Love, Medin, & Gureckis, 2004) would likely work.

In comparison with the large number of well-studied category learning models, relatively few models of function learning have been proposed. *Rule-based models* perform regression given the stimuli (Koh & Meyer, 1991), but are not typically consistent with major experimental results. *Exemplar models* use interpolation and extrapolation techniques to generalize over stored examples (e.g. Bott & Heit, 2004; Busemeyer, Byun, Delosh, & McDaniel, 1997; DeLosh et al., 1997). These models account for many experimentally observed results, but cannot account for multi-modal patterns in interpolation results when discontinuous or externally-cued functions are learned (Lewandowsky, Kalish, & Ngang, 2002). A major step in accounting for that data was taken with the introduction of the Population Of Linear Experts (POLE) model (Kalish et al., 2004). POLE uses a mixture of experts (Jacobs, Jordan, Nowlan, & Hinton, 1991; Erickson & Kruschke, 1998) approach to learning, where a *gating module* learns which of a large number of fixed experts can most accurately make a prediction for a particular stimulus. (See also Harris & Minda, 2005.)

For example, consider a task where a model has to learn to appraise gems based on their caret weight and whether the gem is a diamond or a ruby. A POLE network that has successfully learned this classification will have experts that make accurate predictions for subsets of the space of possible stimuli. One expert might make accurate predictions for diamonds between 0.8 and 1.3 carets, while another might make accurate predictions for rubies between 1.1 and 1.9 carets. The gating module, upon seeing a diamond of 1.1 carets, would assign a large weight to the first expert and a low weight for the second expert.

Although we agree with the mixture of experts approach used in POLE, some of its technical details precluded its use in our work. Whereas POLE, as described by Kalish et al. (2004), requires a large number of fixed linear experts, our new model was designed to use just a few exemplar-based learning experts. Also, Kalish et al. (2004) note that POLE could use experts trained with the Delta rule, but our early experiments with variants of POLE (Harris & Minda, 2005) show that the model's performance using those experts can be very unstable, and tend not to be able to account for the relatively accurate slopes provided by human function learners. Instead, by using exemplar-based experts, a mixture-of-experts model can accurately make predictions when learned exemplars are repeated, and over time can induce stable linear "rules" that can then be selected by the gating module. We suspect that compared to POLE, exemplar-based learning experts will be better able to account for early phases of function learning, and may be more stable overall.

Therefore, we developed a new model of function learning, called the ALCOVE-based Ensemble of Gated Regression Experts (AEGRE). AEGRE uses the off-the-shelf ALCOVE model of categorization (Kruschke, 1992) as a gating module, while its experts use a novel exemplar-based regression technique. See Figure 1. Each expert makes predictions based on only a single continuous dimension[2] (e.g., caret weight). The gating network uses all of the stimulus attributes (caret weight and gem type) as input, and outputs a probability for each expert. The prediction of the AEGRE network as a whole will be the prediction of one of the experts, selected using the probabilities generated by the gating network. Over time, the ALCOVE gating network learns to predict which expert will perform well for each possible stimulus, changing its exemplar weights and attention weights to improve the probabilities used to select the experts. As shown in Figure 1, the gating network's exemplar weights and attention weights are only used internally by the gating module, and do not affect the predictions of the experts.

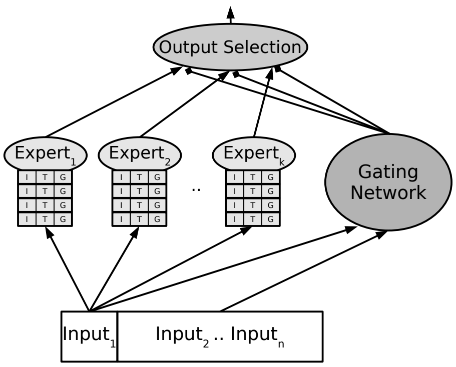

Figure 1: AEGRE function learning system, block diagram. The gating module is an instance of ALCOVE. The experts have associated lists of partial exemplar vectors (with Inputs, Targets, and Gating (weight) values).

Each expert in an AEGRE model maintains a bounded-length list of partial exemplars (just the continuous dimension, not the binary cues), each associated with a stored response and weight. For example, an expert that has learned (with the assistance of the gating network) to specialize in diamonds might have partial exemplars that look like: (1.0 ct, \$2000, 0.8), (1.3 ct, \$1700, 0.7), etc. When predicting a response based on a particular stimulus, the expert performs a weighted least-squares interpolation over its short list of stored partial exemplars, and uses the resulting line as the basis for the prediction. Each expert maintains its list of its x (a parameter) best items, using the probabilities output from the gating network as the measure of "best." If the gating network has successfully learned to partition the input space into approximately linear areas, each expert

[2] Formally, and most generally, function learning can be described as the process of learning the mapping $\Re^n \rightarrow \Re$. However, experimental research into this task has typically been restricted to the mapping $\Re^1 \rightarrow \Re$, from a single continuous stimulus variable to a continuous response, or in some cases to the mapping $\{0,1\}^n, \Re^1 \rightarrow \Re$, from a single continuous stimulus variable, plus any number of additional context (cue) features, to a continuous response. This latter case is the topic of the research presented here.

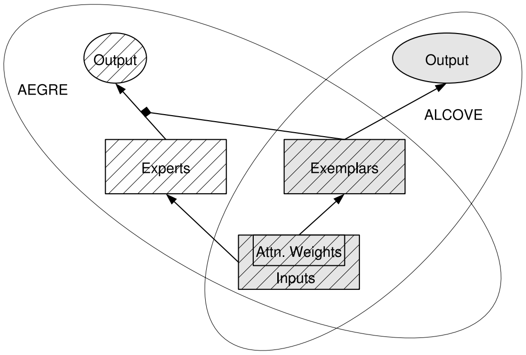

Figure 2: Block diagram of the combined AEGRE and ALCOVE model.

will specialize in exemplars from a particular area of the space of possible stimuli.

Futher details, the mathematical formulation of AEGRE, and implementation issues will be provided in a forthcoming paper (Harris & Minda, submitted).

Combining AEGRE and ALCOVE to form a model capable of accounting for the Minda and Ross (2004) results is straightforward. The combined model consists of the two sub-models running side-by-side, but sharing a single set of attention weights. (Or, equivalently, sharing a single set of attention-weighted exemplar nodes.) See Figure 2 for a schematic diagram of the combined model.

Simulations

Before immediately using the combined model to account for the full set of experimental effects, we first use the AEGRE model to account for the Prediction-Only condition, ignoring shared selective attention and the ALCOVE model until AEGRE's behavior is better understood.

Simulation 1: Prediction-Only

Minda and Ross (2004) reported several measures that we used to fit the model's parameters: mean blocks to convergence and the mean performance on the training data and on the generalization task (compared with CA and FR ideals). We also examined the raw by-subjects results to compute standard deviations and (in experiments not described here) to fit individual subject response patterns.

AEGRE was trained on a rescaled version of the Minda and Ross (2004) task, as follows. The size of the animal was coded as 0.1, 0.5, or 0.9, instead of 1, 2, or 3. Likewise, the target value, the amount of food, was recoded, using a target function of $f = 0.3c + s/2$, where f is the amount of food, c is the category (0 or 1), and s is the size. The model's response was considered correct if it was within 0.05 of the correct response (possible correct responses could differ by no less than 0.1). Rescaling does not affect the results of the simulation, and was done for convenience only.

The model was trained on up to 12 blocks of the training stimuli, following the procedures reported by Minda and Ross (2004). If all 30 responses in a block were deemed correct, training stopped. Following training, the model was run with learning disabled on three test sets: the training set again, the test set with CA responses as the correct responses, and the test set with FR responses as the correct responses. The performance of the model on these three tests is equivalent to the "old items," "conflict items – CA," and "conflict items – FR" that Minda and Ross reported.

Table 2: AEGRE's fit to Minda and Ross (2004), Experiment 1, prediction-only condition. Shown are means and standard deviations of blocks to criterion, accuracy of training items in test phase, and proportion of transfer items responses that match CA-like and FR-like expected responses.

	blocks	training	test = CA	test = FR
MR2004	8.17	.87 (.14)	.62 (.37)	.28 .(37)
AEGRE	8.00	.92 (.22)	.64 (.27)	.31 (.26)

Parameters were tuned using a random-restart hill-climbing method, with χ^2 used to calculate the goodness of fit. (The number of experts was fixed at 2 and was not tuned.) The best results were found with the specificity parameter $c = 2.0$, decision consistency $\phi = 4.35$, exemplars per expert $E = 7$, gating module learning rate $\eta_g = 0.01$, attention learning rate $\eta_\alpha = .64$, and gating module target gain $\tau = 2.97$, giving $\chi^2 = .12$ and the performance shown in Table 2. The table shows the empirical and model fits to the dependent measures.

The final attention weights should also be noted. The AEGRE gating weights cannot be directly compared to the GCM weights reported by Minda and Ross (2004), as AEGRE's attention weights are not constrained to sum to 1 and also include the size dimension. The weights for the criterial attribute averaged 1.35 ($sd = 0.11$), the mean weights for the family resemblance attributes averaged 0.68 ($sd = 0.85$), and mean weight for the size attribute averaged -0.38 ($sd = 0.40$). These values are consistent with AEGRE's gating network selecting experts primarily based on the CA and partially based on the FR values. Note that between different runs of AEGRE, with different initial conditions (analogous to different subjects), the amount of attention to the FR attributes varied by quite a bit. In Minda and Ross's studies, the variance in attention was also quite wide.

Function learning is an inherently more general task than category learning, and function learning models necessarily have more free parameters than do category learning models. AEGRE has seven free parameters, compared with ALCOVE's four parameters. Although the number of free parameters somewhat reduces the power of the fit reported here, the results do allow us to have confidence that AEGRE is a good function learning model to use in parallel concept learning simulations.

Simulation 2: Both Conditions

The experiments of Minda and Ross (2004) involved (for half the subjects) making categorization and scalar pre-

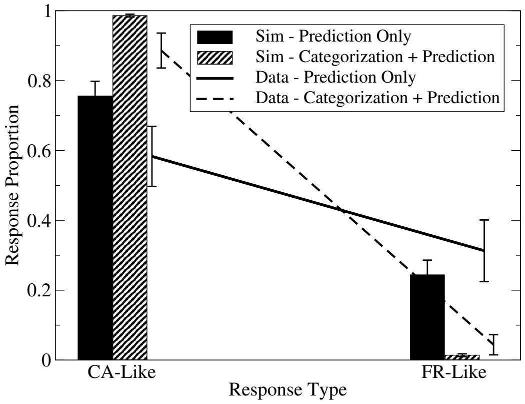

Figure 3: Proportion generalization responses consistent with Criterial Attribute and Family Resemblance patterns, by AEGRE+ALCOVE combined model, compared with data from Minda and Ross (2004). **No** free parameters. Experimental data do not sum to 1.

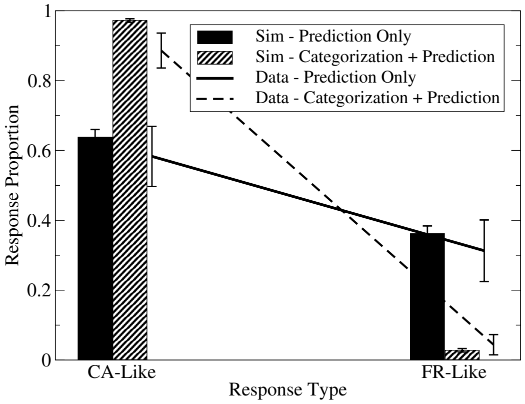

Figure 4: Proportion generalization responses consistent with Criterial Attribute and Family Resemblance patterns, by AEGRE+ALCOVE combined model, compared with data from Minda and Ross (2004). **One** free parameter (η_α). Experimental data do not sum to 1.

dictions based on the same examples in an interleaved manner. To simulate this process, the AEGRE function learning model was run alongside the ALCOVE categorization model. In this new, dual-task model, the attention weights and exemplars of the AEGRE model are also the attention weights and exemplars of the ALCOVE model. On each trial, the ALCOVE parts of the model first are trained to categorize the stimulus, then the AEGRE parts of the model are trained to perform the prediction task. In this manner, both models are updating their shared attention weights using their normal update rules.

The AEGRE model used the parameters identified in Simulation 1 as being best able to fit the aggregate data, while the ALCOVE model used the relevant four parameters (c, ϕ, η_g, and η_α) from AEGRE. Thus, no parameters were additionally tuned for this experiment. The models were run on ten blocks of training data, and then were tested and the results were compared to the CA-like and FR-like expected responses, as in Simulation 1. This process was repeated twice for each paired model, once analogous to the parallel concept learning (Categorization + Prediction) condition, and once analogous to the function learning (Prediction Only) condition. In the latter case, the ALCOVE model was disabled and did not affect the weights or predictions of the AEGRE model. To explore how variance due to the random initial conditions affected the model's performance, the simulations were repeated 30 times, all with identical parameters. After training, no errors in the categorization task were made by any model. On the prediction task, 29 of the 30 models made all predictions consistent with either CA or FR hypotheses, while one model made no consistent predictions; that model was omitted from further analysis. The proportion of responses consistent with CA and FR hypotheses is shown in Figure 3. Note the parallel task effect in the simulation data that closely resembles the same effect in Minda and Ross's experimental data.

The shared attentional weights explain the results. Following training, the mean attentional weights in the Categorization + Prediction condition were consistently 2.0 for the Criterial Attribute, and 0 for the Family Resemblance and Size features. (Recall that attention weights are used by the ALCOVE model and the gating component of the AEGRE model, but not by AEGRE's experts which rely on the Size feature.) This closely reflects the argument made by Minda and Ross (2004), that attention was narrowly distributed in that condition, and more broadly distributed in the Prediction-Only condition (see numbers in Simulation 1).

Note that due to the relatively high attention learning rate found by the parameter optimization in Simulation 1, the model converges to this solution very quickly. This also suggests why the model's responses in the Categorization + Prediction condition are so close to all CA-consistent in Figure 3. It seems that the parameters selected by the search process in Simulation 1, for the Prediction task, result in exceedingly and unrealistically fast learning in the Simulation 2 Categorization task. When the learning rates (η_α) are set (by informal fitting) to 0.1 instead of 0.64, there are much smoother and more realistic changes to average attention weights, and even closer fits to the empirical data. See Figure 4 for the results from the model with only this one parameter roughly tuned.

Discussion

We created a model that begins to account for the interactions among multiple, parallel concept learning tasks. The results support our hypothesis, that shared attention underlies the observed effects. Our simulations used the standard ALCOVE model of categorization and a novel function learning model (AEGRE), and were able to account for important observed data (Minda & Ross, 2004).

The primary contribution of this work is a new framework for thinking about how parallel concept learning tasks interact with each other. We view these tasks as parallel, not superordinate/subordinate, and have proposed that selective attention is shared among the tasks. By sharing attention, parallel concept learning tasks can affect each other's inductive biases. In the Minda and Ross (2004) study, the strong unidimensional bias in the category learning task spreads to the function learning task, which normally seems to have a much weaker unidimensional bias. The subjects' resulting concepts were distinctly different, with different generalization patterns.

Our future work will include further research on AEGRE and how it accounts for function learning experiments, simulations of other multiple-concept learning situations, and experimental tasks to attempt to confirm or disprove our attention hypothesis.

Acknowledgements

Support for this research was provided by NSERC Grant 298512-04 to John Paul Minda, NIMH Grant MH041704 to Gregory L. Murphy and NIDCD Grant DC-005765 to James S. Magnuson. Thanks to Lewis Bott, Gregory Murphy, Jim Magnuson, and others for helpful suggestions, and to Pierre Perruchet, Stephan Lewandowsky, and two anonymous reviewers.

References

Ashby, F. G., & Ell, S. W. (2001). The neurobiology of human category learning. *Trends in Cognitive Sciences*, *5*, 204–210.

Bott, L., & Heit, E. (2004). Nonmonotonic extrapolation in function learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(1), 38–50.

Busemeyer, J. R., Byun, E., Delosh, E. L., & McDaniel, M. A. (1997). Learning functional relations based on experience with input-output pairs by humans and artificial neural networks. In K. Lamberts & D. Shanks (Eds.), *Knowledge, concepts, and categories* (pp. 405–437). Cambridge, MA: MIT Press.

DeLosh, E. L., Busemeyer, J. R., & McDaniel, M. A. (1997). Extrapolation: The sine qua non for abstraction in function learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*, 968–986.

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

Harris, H. D., & Minda, J. P. (2005). Function learning with an ensemble of linear experts and off-the-shelf category-learning models. In *Proceedings of the 27th Annual Conference of the Cognitive Science Society.* Mahwah, NJ: Lawrence Erlbaum Associates.

Harris, H. D., & Minda, J. P. (submitted). Learning two concepts at once: An attentional hypothesis and model.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., & Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural Computation*, *3*, 79–87.

Kalish, M. L., Lewandowsky, S., & Kruschke, J. K. (2004). Population of Linear Experts: Knowledge partitioning and function learning. *Psychological Review*, *111*, 1072–1099.

Koh, K., & Meyer, D. E. (1991). Function learning: Induction of continuous stimulus-response relations. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *17*, 811–836.

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

Lewandowsky, S., Kalish, M., & Ngang, S. K. (2002). Simplified learning in complex situations: Knowledge partitioning in function learning. *Journal of Experimental Psychology: General*, *131*, 163–193.

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

Minda, J. P., & Ross, B. H. (2004). Learning categories by making predictions: An investigation of indirect category learning. *Memory and Cognition*, *32*, 1355–1368.

Murphy, G. L. (2002). *The Big Book of Concepts.* Cambridge, MA: MIT Press.

Nosofsky, R. M. (1992). Exemplars, prototypes and similarity rules. In A. Healy, S. Kosslyn, & R. Shiffrin (Eds.), *From learning theory to connectionist theory: Essays in honor of W. K. Estes* (Vol. 1, pp. 149–168). Hillsdale, NJ: Erlbaum.

Nosofsky, R. M., & Palmeri, T. J. (1996). Learning to classify integral-dimension stimuli. *Psychonomic Bulletin and Review*, *3*, 222-226.

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

Osherson, D. N., Smith, E. E., Wilkie, O., López, A., & Shafir, E. (1990). Category-based induction. *Psychological Review*, *97*, 185–200.

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.

Ross, B. H. (1997). The use of categories affects classification. *Journal of Memory and Language*, *37*, 240–267.

Smith, L. B., Jones, S. S., Landau, B., Gershkoff-Stowe, L., & Samuelson, L. (2002). Object name learning provides on-the-job training for attention. *Psychological Science*, *13*, 13–19.