The Emergence of Multiple Learning Systems

Bradley C. Love (love@psy.utexas.edu)
Matt Jones (mattj@psy.utexas.edu)
Consortium for Cognition and Computation, The University of Texas at Austin
Austin, TX 78712 USA

Abstract

Multiple learning systems models hold that separate learning systems, often organized around discrepant principles, combine their outputs to support human categorization. Rather than propose a complex model, we adopt a complex systems' viewpoint and propose that multiple learning systems emerge from a flexible and adaptive clustering mechanism's interactions with the environment. The model, CLUSTer Error Reduction (CLUSTER), retains the flexibility characteristic of human learning by building knowledge structures as needed to support a learner's goals. Importantly, CLUSTER can apply ostensibly different procedures to different parts of the stimulus space, a hallmark of multiple systems models. We describe a simulation of a human learning study in which CLUSTER develops different cluster representations for different item types. Rule-following items are captured by clusters that are broadly tuned and focused on rule-relevant stimulus aspects, whereas exceptions (especially those that violate high-frequency rules) are captured by narrowly tuned clusters that focus on item-specific stimulus qualities. We end by considering the relation between CLUSTER and findings from the cognitive neuroscience of category learning.

Introduction

Proposals for category representation are diverse, ranging from exemplar- (Medin & Schaffer, 1978) to prototype-based (Smith & Minda, 1998) and include proposals between these two extremes (Love, Medin, & Gureckis, 2004). Determining the best psychological model can be difficult as one model may perform well in one situation but be bested by a competing model in a different situation. One possibility is that there is not a single "true" model.

In category learning, this line of reasoning has led to the development of models containing multiple learning systems. These more complex models hold that category learning behavior reflects the contributions of different systems organized around discrepant principles that utilize qualitatively distinct representations. The idea that multiple learning systems support category learning behavior enjoys widespread support in the cognitive neuroscience of category learning (see Ashby and O'Brien, 2005, for a review and Nosofsky and Zaki, 1998, for a dissenting opinion).

Multiple system models of category learning detail the relative contributions of the component learning systems. The relative contributions can depend on the circumstances. For example, ATRIUM (Erickson & Kruschke, 1998) contains a rule and exemplar learning system. Which system is operable is determined by a gating system, allowing different classification procedures to be applied to different parts of the stimulus space. For example, familiar items could be classified by the exemplar system whereas rules could be applied to unfamiliar items. The power to apply qualitatively different procedures to different stimuli is the hallmark of multiple systems models.

Proposing multiple systems begs the questions of how many systems are present and how do they interact. Are there two, three, or thirty-four systems? Do some systems combine outputs whereas others shunt each other? These questions are not trivial to answer. For example, a two system model may suffice for one data set, but a new manipulation could provide evidence for a third system. As systems propagate, the complexity of the overall system dramatically increases. Building in this degree of complexity complicates model evaluation.

Instead of proposing a complex model of category learning containing multiple systems, we advocate a complex systems approach to category learning modeling in which multiple learning systems emerge from a flexible and adaptive clustering mechanism's interactions with the environment. We evaluate the hypothesis that a relatively small set of learning principles can effectively "grow" knowledge structures that satisfy the needs that multiple systems models are intended to address.

Past Work and Current Challenges

Previous work with the SUSTAIN model, which is the precursor to the model that we introduce here, has partially delivered on the promise of flexibly building needed knowledge structures. SUSTAIN is a clustering model that starts simple and recruits clusters in response to surprising events, such as encountering an unfamiliar stimulus in unsupervised learning or making an error in supervised learning. Surprising events are indicative that the existing clusters do not satisfy the learner's current goals and that the model should grow new knowledge structures (i.e., clusters). These clusters are modified by learning rules that adjust their position to center them amidst their members. Dimension-wide attention is also adjusted to accentuate stimulus properties that are most predictive across clusters.

Although simple, these growth dynamics allow SUS-

TAIN to address a wide range of human learning data across various paradigms including unsupervised, inference, and classification learning (Love et al., 2004). Depending on the circumstances of the learning situation (i.e,, depending on what the task stresses and target categories), SUSTAIN can evolve clusters that resemble prototypes, exemplars, or rules (Love, 2005). Careful behavioral experiments support the conclusion that SUSTAIN is not merely mimicking these other models, but that human learners' and SUSTAIN's representations are in accord (Sakamoto & Love, 2004). In summary, SUSTAIN accounts for both classical studies of category learning and the more contemporary work that suggests that conceptual organization is determined by the interplay of information structures in the environment and task pressures or goals (Markman & Ross, 2003).

Despite these successes, considerable challenges remain. One basic challenge is to formalize the notion of a goal or task pressure and tie the formalism to cluster development. SUSTAIN can capture basic distinctions, such as that between unsupervised and supervised learning, but more can be done. Ideally, the notion of goal would be more encompassing and continuous to capture all possible cases from pure classification learning in which the only goal is to predict category membership to pure unsupervised learning in which the goal is to predict every feature (i.e., to capture the correlational structure of the environment in an unbiased fashion). Importantly, the formal notion of goal should directly affect the recruitment and modification of clusters in a principled way. Learning rules should update clusters to reflect the goal measure and clusters should be recruited in light of how well the current clusters satisfy the current goal measure.

A second basic challenge for adaptive clustering approaches is to display the flexibility necessary to develop clusters that approach the range and richness of the representations that human learners build when learning from examples. For instance, SUSTAIN's attentional mechanism accentuates certain features that are predictive in the current task, but is constrained such that every cluster is focused on the same set of properties. In contrast, people stress different properties in different domains. For example, when shopping for clothing, color is important, but when shopping for a computer the type of processor is important (a feature not even relevant to clothing). To evolve these kinds of knowledge structures and to apply different "procedures" to different parts of the stimulus space as multiple systems models do, each cluster needs to be able to accentuate the features that satisfy the learning goals for the stimulus aspects it represents.

Another instance in which future clustering models need to display greater flexibility is in developing representations at different levels of specificity. Human learners develop concepts that range from very specific (e.g., Jim's dog Fido) to very broad (e.g, living things). Simultaneously capturing regularities at different scales is a formidable challenge. Further complicating matters is that regularities at narrower scales can be embedded within regularities at broader scales. For instance, exceptions to patterns fall within the scope of the overall pattern and stimuli can be categorized at multiple category levels. To address these issues, clusters need to fine tune their level of specificity to satisfy the goal measure. As in the case of adjusting attention at the individual cluster level, adjusting specificity at the individual cluster level allows for different criteria to be applied to different parts of the stimulus space, as in multiple systems models.

The model that is introduced in the next section, CLUSTer Error Reduction (CLUSTER), meets these stated challenges. CLUSTER incorporates a formal goal measure that directs cluster development. CLUSTER has sufficient flexibility to evolve conceptual structures (i.e., clusters) that reflect key aspects of human knowledge representation. After introducing the model, the formalism will be presented and a supportive simulation will be discussed. The simulation illustrates how CLUSTER can evolve cluster organizations that serve the functions of multiple systems. Finally, we will consider how CLUSTER is consistent with cognitive neuroscience findings advocating multiple memory systems and briefly discuss work that is being done to further develop and verify the model.

Overview of CLUSTER

CLUSTER is an auto-associative model of human category learning in which the "hidden" layer consists of clusters (see Figure 1). A presented stimulus activates the existing clusters, which pass their activation to the output layer via connection weights. Like other auto-associative models (e.g., Kurtz, 2004), CLUSTER attempts to replicate the input layer at the output layer and in the process develops internal representations that seize on key regularities.

CLUSTER differs from other auto-associative models in a critical way. The error term CLUSTER minimizes does not uniformly weight reconstruction error equally across features. Instead, each feature's error is weighted according to its goal relevance. For example, pure classification learning places all the error term weighting on the category label features and error associated with reconstructed perceptual features is disregarded (as in most category learning models). At the other extreme, pure unsupervised learning weights the reconstruction error uniformly across features (as in most auto-associative models). CLUSTER can capture every conceivable case in between these extremes, which is critical as the extremes are likely cartoons that do not correspond to human learning (e.g., people incidentally learn about feature correlations in classification learning and place more importance on predicting certain features in unsupervised learning).

The error term (with goal weights on each feature) is the learning goal, formally stated. To satisfy this goal, clusters adjust their position, attention, and weights to minimize the error term through gradient descent learning. Thus, depending on the goal weights, different cluster organizations will emerge. Unlike most models,

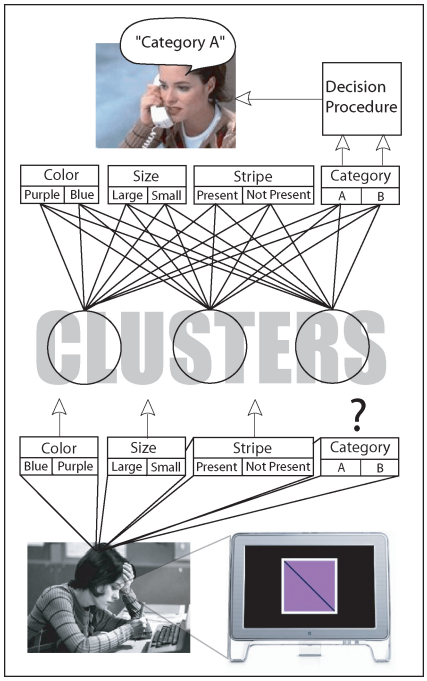

Figure 1: CLUSTER is an auto-associative learning model in which the hidden layer consists of clusters that adjust their position, attention, and association weights to minimize an error term that reflects the learner's goals. In the illustrated example, three clusters have been recruited and the model is being asked to infer the category label.

each cluster can adjust its own attention to minimize error and attention does not sum to a fixed number. These changes allow additional flexibility for clusters to emphasize different features and to vary in specificity (e.g., a specific dog vs. dogs in general). Although Figure 1's grouping of features implies dimensional structure, CLUSTER departs from the majority of models that utilize selective attention mechanisms (e.g., Nosofsky, 1986) in that it does not assume a dimensional structure. Not assuming dimensional structure allows for additional flexibility (e.g., the presence or absence of red can be critical to a cluster, whereas the presence or absence of blue can be somewhat irrelevant).[1]

CLSUTER begins with one cluster centered on the first training example and recruits additional clusters when the existing clusters are not supportive of the current goal. Each newly recruited cluster is centered upon the current stimulus. Like CLUSTER's other operations, the algorithm for cluster recruitment is consistent across all induction tasks (there are no special cases). Despite its consistency across situations, CLUSTER retains the flexibility to build representations that capture many of the competencies of human learners without proposing distinct learning systems. CLUSTER is highly principled (all of its operations are tied to the goal-weighted error term), but minimal structure is built in to the model. Instead, CLUSTER evolves the knowledge structures needed to solve the current task.

[1]Interestingly, in cases in which contrasts are consistent (e.g., when red is present, blue is absent, and vice versa), CLUSTER attends equally to the contrasting features within each cluster. Thus, CLUSTER may prove to provide some insight into how dimensional structure arises.

CLUSTER's Formalism

This section presents the equations that define CLUSTER. First, we consider how CLUSTER generates a response. Then, we consider how CLUSTER learns.

CLUSTER: Generating a prediction The distance between the stimulus and each cluster is calculated. The attention-weighted distance I^j between the all the known features of stimulus S and cluster j is:

$$I^j = \sum_{i=1}^{m} \alpha_i^j (H_i^j - S_i)^2 \quad (1)$$

where m is the number of stimulus features, α_i^j is cluster j's attentional weighting of feature i, H_i^j is cluster j's position along feature i (i.e., the value cluster j expects along feature i). Each S_i is 0 (absent) or 1 (present) for discrete features and ranges between 0 and 1 for continuous features. Unknown features are ignored when calculating distance.

The output of cluster j is:

$$A^j = \lambda_j \cdot e^{-I^j} \quad (2)$$

where λ_j is the sum of cluster j's attentional weighting for all known features. One subtle difference with most models is that the receptive field function for clusters is Gaussian instead of exponential. This functional form allows for peak-shift responding in which stimuli outside the experienced range of examples can lead to responses more deterministic than for experienced stimuli, consistent with rule-based generalization behavior to unfamiliar stimuli.

Activation passes from the clusters to the output units via association weights. The output of unit i is:

$$O_i = \sum_{j=1}^{n} w_j^i \cdot A^j + .5 \quad (3)$$

where w_j^i is the association weight from cluster j to output unit i. Outputs are truncated to lie between 0 and 1. The default value of .5 can be conceived of as a prior over features.

In discrete-feature prediction tasks in which one of a set of unknown features must be chosen (e.g., predicting the category label in a classification task), the probability of choosing unknown feature k is:

$$Pr(k) = \frac{(O_k)^d}{\sum_{l=1}^{v} (O_l)^d} \quad (4)$$

where d is a decision parameter, and l ranges over the v features forming the choice set. The power response

rule is chosen over an exponential form to enable the aforementioned peak-shift responding behavior.

In recognition tasks, the recognition strength for stimulus S is given by the sum of all cluster activations resulting from the presentation of S.

CLUSTER: Learning and Cluster Recruitment After feedback, full stimulus information is known. Gradient descent learning minimizes the error between the stimulus and CLUSTER's reconstruction of it at the output layer:

$$E = \frac{1}{2}\sum_{i=1}^{m} \delta_i \cdot (S_i - O_i)^2 \tag{5}$$

where δ_i is the goal weighting for feature i subject to the following constraints:

$$\sum_{i=1}^{m} \delta_i = 1 \text{ and } \forall i \ \delta_i \geq 0. \tag{6}$$

Gradient descent learning rules minimize error by adjusting each cluster's position, attention, and association weights. These learning rules are derived by differentiating the error term with respect to the adjusted quantity (i.e., position, attention, association weights). Each learning rule has an associated learning rate parameter.

After receiving feedback but prior to applying the learning rules, new clusters are recruited when the existing clusters are a poor match to the current stimulus. Specifically, a new cluster is recruited when:

$$\frac{\sum_{j=1}^{n} A^j \cdot G^j}{\sum_{j=1}^{n} A^j} < \tau \tag{7}$$

where τ is the recruitment threshold parameter and cluster j's goodness is:

$$G^j = 1 - \sum_{i=1}^{m} \delta_i \cdot (H_i^j - S_i)^2. \tag{8}$$

A newly recruited cluster is centered on the current stimulus item. Association weights are set to zero. The new cluster's sum of attention for the features known at the initial stimulus presentation is set to be p (a parameter) above the value ξ necessary to prevent recruitment if the current stimulus was re-presented:

$$\lambda^{n+1} = \xi + p \tag{9}$$

where

$$\xi = \frac{\tau \cdot \sum_{j=1}^{n} A^j - \sum_{j=1}^{n} A^j \cdot G^j}{1 - \tau}. \tag{10}$$

Attention is allocated uniformly to all features (known and unknown) and is set to λ^{n+1} divided by the number of known features at initial stimulus presentation.

Illustrative Simulation

Findings from previous studies exploring rule-plus-exception learning have been problematic for exemplar models and have been used to support multiple systems models, like the RULEX model of category learning (Nosofsky et al., 1994). RULEX proposes that rule-following items are captured by a rule system whereas exception items reside in an exemplar store. Here, we demonstrate that CLUSTER can accommodate such findings by applying different procedures to different parts of the stimulus space and in fact provides an account superior to RULEX's.

To test between this dual route account (i.e., rules and exceptions) and a clustering account, Sakamoto and Love (2004) revisited the rule-plus-exception design with the twist that one rule was twice as frequent as the other. Subjects sequentially classified stimulus items into categories A and B and received corrective feedback. Each category was defined by a rule (e.g., if large, then A; if small, then B). Additionally, each category contained an exception (e.g., a small member of A; a large member of B). Table 1 provides the design details of Sakamoto and Love's variation in which one experienced category had twice as many rule-following items as the contrasting category. Following learning, recognition memory was assessed. In contrast to RULEX's predictions (across all explored parameter values), the exception violating the more frequent rule was better remembered than the exception violating the less frequent rule (see Figure 2).

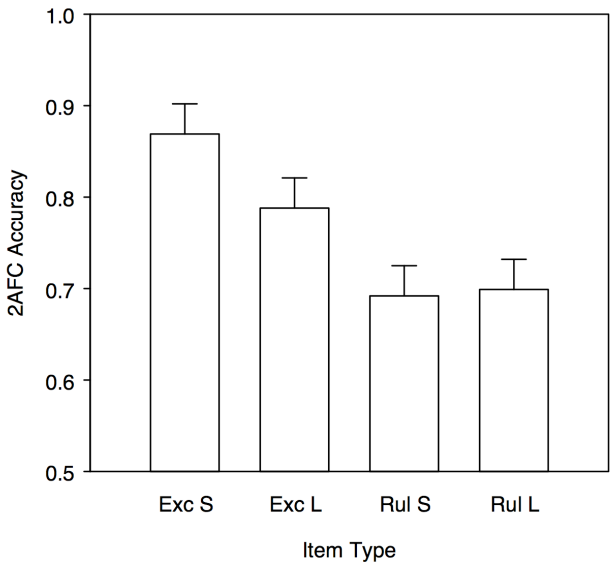

Figure 2: Mean accuracies in the recognition phase of Sakamoto and Love's (2004) Experiment 1 are shown along with 95% within-subjects confidence intervals (see Loftus & Masson, 1994). Exc S is the exception of the small category, Exc L is the exception of the large category, Rul S are the rule-following items of the small category, and Rul L are the rule-following items of the large category.

CLUSTER was applied to the data to illustrate its ability to "evolve" multiple systems. Each stimulus dimension shown in Table 1 and category membership were

Table 1: The abstract stimulus structure for Sakamoto and Love's (2004) Experiment 1 is shown. Items A1 and B1 (indicated by the arrows) violate the imperfect rule of the first stimulus dimension. Subjects completed 10 training blocks where each block consisted of each item below presented in a random order. Following learning, Items A1-5 and B1-B5 were paired with all combinations of novel foils that matched on the first dimension in forced choice recognition. The actual stimuli were simple geometric figures. For example, for some subjects the first dimension was size with a 1 indicating a small figure and 2 indicating a large figure.

Learning Items	Dimension Values	Novel Items	Dimension Values
Category A			
→ A1	21112	N1	11221
A2	12122	N2	12112
A3	11211	N3	12221
A4	12211	N4	12212
A5	11122	N5	12222
A6	12111	N6	21221
A7	11222	N7	22112
A8	11212	N8	22221
A9	12121	N9	22212
Category B		N10	22222
→ B1	11121		
B2	22122		
B3	21211		
B4	22211		
B5	21122		

represented by 2 features for a total of 12 features. In contrast to RULEX (which requires eight parameters to CLUSTER's seven for the simulation), multiple sets of parameters replicated the basic pattern of results, indicating that these findings follow from CLUSTER's basic operation and that additional work is necessary to establish default parameters for CLUSTER. These and other model evaluation issues, such as consideration of nested models within CLUSTER's formalism, are topics currently being intensely pursued, but are set aside here in favor of demonstrating CLUSTER's promise to evolve multiple learning systems. In this spirit, the following parameters were selected because of the interpretability of the resulting simulations: $\tau = .3250$, $p = 8.5$, and the learning rates for attention, position, and weights were .001, 10.0, and .1 respectively. The δ values were set such that .9 of the total sum of 1 was devoted to the category label features with the remaining features weighted uniformly (i.e., the model's primary goal was to correctly classify the stimulus, but some importance was given to learning about relationships predicting other features). Finally, because the rule dimension (i.e., the first dimension) was cued for subjects, this dimension was made more salient by allocating 91% of attention to the two features forming this dimension when a new cluster was recruited. Because the learning data is not discussed here (CLUSTER does the fit the pattern), the d parameter was not used. In forced-choice recognition, the item with the higher recognition strength was taken as CLUSTER's choice.

Using these parameters, CLUSTER was simulated 10,000 times adopting methods paralleling the human study (e.g., 10 blocks of training) and the results were averaged. Adopting the labels from Figure 2, CLUSTER predicts Exc S=.88, Exc L=.80, Rul S=.58, and Rul L=.59, which replicates the two major findings: exceptions are better remembered than rule-following items with the exception violating the more frequent rule (i.e., the exception in the small category) being best recognized.

CLUSTER recruited 11.4 clusters on average (the median was 11) to represent the 14 training items. The number of clusters recruited followed a normal distribution with solutions ranging from 4 to 23 clusters with a standard deviation of 2.3. Every solution examined involved devoting at least one cluster to encoding each exception with many simulations devoting multiple clusters to each exception. Because CLUSTER is a distributed model and its predictions for an item depend on the responses of all clusters, an analysis of the four item types was conducted that factored in all clusters.

One explanation for CLUSTER's ability to accomodate the results is that it increased attention for clusters playing prominent roles in coding the exceptions, particularly for non-rule stimulus features. Encoding these items at a different specificity than rule-following items would help reduce confusions between these items and rule-following items, resulting in both reduced error during training and in enhanced recognition for exceptions. The pressure to enhance attention should be greatest for the exception violating the more frequent rule as every rule-following item from the contrasting category provides an impetus to enhance attention.

To evaluate this explanation, following training, study items were presented to CLUSTER and a weighted sum of attention to non-rule features was calculated by multiplying each cluster's sum of attention for non-rule features by its activation. Then, these products were summed and normalized by dividing by the sum of all cluster activations. The results for items of the same type were averaged. The mean results for the four item types (averaged over 10,000 simulations) are Exc S=1.36, Exc L=1.32, Rul S=1.28, and Rul L=1.29. As predicted, these sums perfectly track item recognition. Exceptions (particularly the exception violating the more frequent rule) were stored as "hot spots" of focused activity whereas clusters coding for rule items were more broadly tuned and were less apt to code item specific differences. Distinct representations emerge for the item types.

Discussion

Human learners display flexibility in how they represent category information that outstrips the capacities of tra-

ditional single system models. In response, the field has developed multiple system models that are themselves not without problems. Here, we pursue a third approach – knowledge structures evolve as needed to satisfy the learner's goals.

CLUSTER embodies this third position. CLUSTER has a formally defined notion of goal that spans inductions tasks, recruits clusters when existing clusters fail to support the learner's goals, and adjusts clusters' positions, attention, and association weights to reduce goal mismatch. These operations are sufficient to apply different procedures to different parts of the stimulus space, as multiple systems models do.

How do we reconcile our position with impressive evidence from cognitive neuroscience that multiple systems underly human category learning performance? We do not deny that multiple learning systems underly human category learning. A non-exhaustive list of systems includes a dopaminergic procedural learning system, working memory system engaging cortical-thalamic loops, and a PFC-hippocampal-perirhinal learning system. The latter system is marked by its flexibility and is adept at creating new conjunctive representations that link features (i.e., clusters). SUSTAIN (the precursor to CLUSTER) has been put in alignment with this learning circuit and has successfully simulated populations with hippocampal deficits (Love & Gureckis, under review). CLUSTER likely corresponds to the hippocampal system as well. We believe that a fast learning hippocampal system is shadowing the other learning systems. For instance, the literature is replete (including Sakamoto and Love, 2004) with cases in which learners are clearly applying a rule stored in working memory, but are nevertheless storing additional information about rule-following examples. Another way to reconcile CLUSTER with a multiple learning systems view is to view these systems emerging over an evolutionary time scale.

Much work remains to be done. Efforts are underway to apply CLUSTER to all the studies to which SUSTAIN has been applied. The results so far are promising. Additionally, we are applying CLUSTER to studies exploring how people partition knowledge and appear to apply different procedures depending on context (e.g., Yang & Lewandowsky, 2004).

Finally, CLUSTER has been successfully applied to Kruschke's (1993) filtration and condensation tasks that were intended to demonstrate the necessity of dimensional attention (CLUSTER has cluster and feature specific attention). Although CLUSTER does not have a built in notion of dimensional attention, dimensional attention emerges (i.e., there is advantage for aligning all clusters along the same contrasting features) much like how what looks like multiple learning systems emerges out of the Sakamoto and Love (2004) simulations. While CLUSTER itself is still evolving, it appears it has the necessarily constraints built in to account for human learning and no more.

Acknowledgments

This work was supported by AFOSR grant FA9550-04-1-0226 and NSF CAREER grant 0349101 to B. C. Love and NIH NRSA F32-MH068965 to M. Jones.

References

Ashby, F., & O'Brien, J. B. (2005). Category learning and multiple memory systems. *Trends in Cognitive Sciences*, *9*, 83-89.

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

Kruschke, J. K. (1993). Human category learning: Implications for backpropagation models. *Connection Science*, 3-36.

Kurtz, K. J. (2004). The divergent autoencoder (DIVA) account of human category learning. *Proceedings of the Cognitive Science Society*, 1214-1219.

Loftus, G. R., & Masson, M. E. J. (1994). Using confidence intervals in within-subject designs. *Psychonomic Bulletin & Review*, *1*, 476–490.

Love, B. C. (2005). Environment and goals jointly direct category acquisition. *Current Directions in Psychological Science*, *14*, 195-199.

Love, B. C., & Gureckis, T. M. (under review). Models in search of a brain.

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

Markman, A. B., & Ross, B. H. (2003). Category use and category learning. *Psychological Bulletin*, *129*, 592-613.

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

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

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review*, *101*(1), 53-79.

Nosofsky, R. M., & Zaki, S. F. (1998). Dissociations between categorization and recognition in amnesic and normal individuals. *Psychological Science*, *9*, 247-255.

Sakamoto, Y., & Love, B. C. (2004). Schematic influences on category learning and recognition memory. *Journal of Experimental Psychology: General*, *33*, 534-553.

Smith, J. D., & Minda, J. P. (1998). Prototypes in the mist: The early epochs of category learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *24*, 1411–1430.

Yang, L. X., & Lewandowsky, S. (2004). Context-gated knowledge partitioning in categorization. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *30*, 1045-1064.