Some Attention Learning "Biases" in Adaptive Network Models of Categorization

Toshihiko Matsuka Chiba University
James E. Corter Teachers College, Columbia University
Arthur B. Markman The University of Texas

Abstract

In two simulation studies, we compare the attention learning predictions of three well-known adaptive network models of category learning: ALCOVE, RASHNL, and SUSTAIN. The simulation studies use novel stimulus structures designed to explore the effects of predictor diagnosticity and independence, and differentiate the models regarding their tendencies to learn simple rules versus exemplar-based representations for categories. An interesting phenomenon is described in which the models (especially SUSTAIN and RASHNL) learn to attend to a completely nondiagnostic constant dimension.

Keywords: category learning; selective attention; simulation.

Introduction

A key assumption of many computational models of categorization is that category learners do not merely form associations between instances and categories, but also learn how to allocate attention to each individual stimulus "dimension" (e.g., color). The present paper focuses on three such adaptive network models of classification learning: the ALCOVE model of Kruschke (1992); RASHNL (Johansen & Kruschke, 1999); and SUSTAIN (Love & Medin, 1998). These models are multilayer adaptive network models that accept as input a stimulus description (in the form of a set of input feature values), and produce as output category membership predictions that are based on the activation levels of a set of output nodes that correspond to the possible category responses. Over the course of training, these models learn both what dimensions to attend to, and how to correctly classify all the stimuli in the training set.

These three adaptive network models differ in several key aspects. ALCOVE and RASHNL are exemplar models, in the sense that each stimulus in the training set is allocated a node in the "hidden" or middle layer of the network. In contrast, SUSTAIN can form either exemplar-level or prototype-based representations. Prototypes are handled by using a reduced number of nodes in the hidden layer, corresponding to potential generalizations. SUSTAIN dynamically allocates new prototypes, allowing it to possibly use *multiple* prototype nodes for each category defined by the training feedback.

Exploring how these models adapt their attention weights is crucial to understanding their usefulness and validity by relating their learning accuracy predictions more directly to learning strategies. In previous studies (e.g., Matsuka & Corter, & Markman, 2002; Corter, Matsuka, & Markman, 2007), we found that all three models can account for human classification accuracy learning curves, but show distinct patterns in their "learning curves" for dimensional attention weights. In particular, ALCOVE and RASHNL seem to pay more attention to relatively independent predictors, while SUSTAIN shows the reverse pattern. The present Simulation 1 seeks to confirm this finding with a novel stimulus structure designed for this purpose, while Simulation 2 investigates an interesting phenomenon whereby the models sometimes learn to pay attention to a completely nondiagnostic feature. First, we briefly describe the models.

ALCOVE (Kruschke, 1992) is a multi-layer adaptive network model of categorization based on the Generalized Context Model (Nosofsky, 1986). The first layer of ALCOVE is a stimulus input layer. Each node in this layer represents the value of the presented stimulus on a single dimension. Importantly, each dimension has an attention strength (α_i) associated with it. Typically, attention strengths are initially equal across dimensions. However, the model learns to reallocate attention as learning proceeds, by adjusting these weights. The second layer in the network is the exemplar layer. Each node in this layer corresponds to an exemplar, described by its position in the multidimensional stimulus space. The activity of the exemplar nodes is fed forward to the third layer, the category layer, whose nodes correspond to the categories being learned. Separate learning rates are assumed for the association weights and attention strengths.

RASHNL (Kruschke and Johansen, 1999) is a modified and extended version of ALCOVE. The modifications introduced in RASHNL include: limited attention capacity; a capability for large and rapid shifts of attention; a gradually decreasing learning rate; and a parameter for salience of cues or features. RASHNL's architecture is similar to that of ALCOVE. However, each dimension has a dimensional salience parameter, the values of which are prespecified by the experimenter (i.e., not adjusted by learning). The dimensional attention strengths, α_i, are derived functions of separate underlying parameters, termed the "gains", which are adjusted by learning. An additional parameter P is incorporated, that can be set to vary between fixed attention capacity ($P = 1$) or unlimited attention capacity ($P = \infty$).

SUSTAIN (Love & Medin, 1998; Love, Medin & Gureckis, 2004), is comprised of two separate adaptive network components, a "supervised" network and an "unsupervised" one. The unsupervised network is a competitive network that clusters stimuli into prototypes. The term 'prototype' is used broadly, however, because an experimenter-defined category might be represented by one or many prototypes, and a prototype might represent only a single stimulus. This

flexibility also gives SUSTAIN the capability to form prototype-plus-exception representations or even exemplar-level representations. This clustering network is dynamic and incremental in its behavior, in the sense that new prototypes and/or exceptions are created when current prototypes are not predictive.

The "supervised" network is a feedforward network that classifies a stimulus based on similarity between the input pattern and the prototypes created by the unsupervised network. The activation of node *j* in the internal layer depends on several parameters: λ_i, which represents the "tuning" of the receptive field for a given dimension *i*, the distance between the centroid of prototype unit *j* and the input node on dimension *i*, and *r*, an overall attentional parameter that can be adjusted to create tighter or looser focus on highly tuned dimensions. The "tuning" (λ_i) parameters in SUSTAIN are the primary determinants of differences in attention among dimensions. When λ_i is large, difference between the input and the prototype node on dimension *i* are "stretched" or emphasized. At the output layer, SUSTAIN allows only the internal-layer unit with the highest post-transformed activation to determine output node activations, leading to "winner-take-all" learning.

Comparing the Models' Accounts of Attention Learning

We are interested in the attention learning behavior of these models. One clear difference between models is that RASHNL was designed with multiple attention learning iterations on each trial, in order to account for rapid shifts in attention that ALCOVE cannot predict. However, other differences among the models' assumed attention mechanisms have unknown implications. For example, it is not clear what follows from SUSTAIN's use of feedback from only the most-activated prototype to update the dimensional tuning parameters. Because of the complexity of these multilayer network models and their dynamic nonlinear performance, simulation studies are useful to establish the models' actual attention-learning behavior in complex learning tasks.

SIMULATION STUDIES

Simulation 1

Our previous findings (e.g., Corter et al., 2008; Matsuka et al. 2002) suggest that ALCOVE and RASHNL tend to incorporate dimensions that are relatively independent, even orthogonal, to the other predictors, compared to SUSTAIN. As an alternative (but related) hypothesis, it may be that relatively independent predictors are preferred by ALCOVE and RASHNL because such dimensions often are more useful for distinguishing exemplars, especially between categories. Simulation 1 explores this hypothesis by decoupling predictor diagnosticity (correlation with the criterion), predictor independence (inversely related to correlation with the other predictors), and "exemplar separation" (i.e., whether a predictor can be used in conjunction with other strong predictors in order to distinguish exemplars from different categories).

Method: Table 1 shows the category structure used in Simulation 1. In a typical classification learning task the classes (A and B) might be diseases, the exemplars patients, and the five "dimensions" might represent five types of test results or symptoms (each with two possible values). Correlations with the criterion are equal to .6 for Dimensions D1 and D2, to .2 for D3 and D4, and zero for D5. D3 and D4 differ in their configural validities, however: The variable subset (D1, D2, D3) gives a perfect R-square (RSQ) of 1.0 when these three dimensions are used in a linear model predicting the criterion, while the variable subset (D1, D2, D4) yields an RSQ of only .77. Addition of the orthogonal variable D5 alone does not increase the RSQ of the predictor set (D1, D2), which is equal to .60.

The dimensions also differ in their degree of independence from the other predictors. Dimension D3 is correlated .6 with D1 and with D2, while D4 is correlated -.2 with each of these two predictors. D5 has a zero correlation with all the other predictors and the criterion. However, the predictors D3-D5 are all comparable in one regard: each one can be used in conjunction with D1 and D2 to distinguish all category A exemplars from all category B exemplars. Thus, the simulation results for this structure should shed light on our hypothesis that this "exemplar separation" measure is key to predicting ALCOVE's and RASHNL's attention allocation behavior, by holding this factor constant across the "extra" dimensions D3-D5.

Table 1. Stimulus structure used in Simulation 1.

Class	D1	D2	D3	D4	D5
A	1	1	1	1	1
A	1	1	1	0	0
A	1	1	1	0	1
A	1	0	0	1	0
A	0	1	0	1	1
B	1	0	1	0	1
B	0	1	1	0	0
B	0	0	0	1	1
B	0	0	0	1	0
B	0	0	0	0	1

Using the three models, we simulated subjects (N=10,000) who were trained for 20 blocks on the stimulus structure shown in Table 1. For each individual subject parameter values were randomly selected from a uniform distribution within reasonable limits for each parameter. The main results recorded were the final-block attention weights for the five dimensions.

Results: Although we cannot identify any of the simulated subjects as being descriptively more plausible than others due to the lack of empirical data for this structure, we can assess the normative success of each simulated subject, by calculating their predicted final-block classification accuracy. Table 2 shows the mean final-block attention parameters for each model, by dimension. The table shows the final weights only for "successful" simulated learners, those achieving at least 80% correct classification accuracy

by the final block. The results do not differ if all simulated learners are included, however. All three models give highest attention weight to the two high-diagnostic dimensions D1 and D2. However, they differ widely in how they distribute attention to the three remaining dimensions. In particular, the results for ALCOVE show a surprising pattern, with nearly as much attention paid to D4 and D5 as to the two most diagnostic dimensions and with D3 weighted least, even though D3 has the highest configural validity (RSQ = 1.0) in conjunction with D1 and D2. Thus, this pattern of weights can be said to be non-optimal; it is a surprising result in that D5 is completely uncorrelated with the criterion. This ordering is consistent with the hypothesis that ALCOVE prefers relatively independent predictors, and cannot be ascribed to differences in "exemplar separability", because this latter factor is held constant for D3, D4 and D5.

Table 2. Simulation 1: Final block relative attention weights for dimensions for each model, for "successful" simulated learners, with number (N) of successful learners out of 10,000 total.

	N	D1	D2	D3	D4	D5
ALCOVE	8480	.248	.247	.098	.199	.209
RASHNL	7463	.274	.286	.183	.123	.135
SUSTAIN	6855	.240	.248	.230	.136	.148

RASHNL and SUSTAIN both predict normatively satisfactory patterns of attention weights in the sense that they give highest attention to D1 and D2, with D3 third highest. This set of predictors is the minimal sufficient set for perfect prediction, therefore these weights may be considered to be the monotonically "optimal" weights. However, both RASHNL and SUSTAIN weight D5 higher than D4. Again this is surprising, since D5 has zero correlations with the criterion (but also with the other predictors).

Discussion: In this simulation RASHNL and SUSTAIN yielded weights that are normatively justifiable by the customary criterion of "configural validity", by giving highest weighting to the three dimensions yielding a perfect multiple-R in predicting the criterion. However, they still gave nontrivial weights to the two remaining dimensions, D4 and D5. In this sense their attention allocation patterns cannot be described as optimal. Furthermore, most simulated learners gave attention to more than one of these "supplementary" dimensions, showing that the network models do not always learn minimal sufficient rules.

ALCOVE also gave highest weights to D1 and D2, but gave third highest weight to D5, a dimension that has a correlation of zero with the criterion and with all the other predictors. This pattern seems "irrational' by the usual criterion of configural validity. However, we note that it is reasonable from the standpoint of "exemplar separability": by this measure, the set (D1, D2, D5) is adequate for the classification task. ALCOVE also gives non-trivial weights to the remaining two dimensions, D3 and D4, again demonstrating that the network models do not tend to learn minimal representations across a broad range of parameter values. Finally, ALCOVE weights D5 higher than D4 and D4 higher than D3, an ordering that is consistent with the degree of independence of the three dimensions, while RASHNL weights D5 over D4 (but weights D3 highest, in line with its configural validity). This result supports the hypothesis that ALCOVE tends to give higher weight to more independent dimensions, even at the cost of finding a non-optimal solution. RASHNL and SUSTAIN both find the "optimal" configuration of dimensions (D1, D2, D3), and in fact exhibit the same ordering of weights (D1≈D2>D3>D5>D4). However, given that only SUSTAIN weights D3 nearly as high as the two diagnostic dimensions, the results are consistent with the hypothesis that this model "prefers" dimensions that are correlated with other important predictors, compared to the other models.

Simulation 2

Simulation 2 explores two issues. The first is the idea that SUSTAIN favors dimensions that are correlated with other predictors, at least relative to the other models. The second issue is the tendencies of the models to utilize exemplar versus simple rule based strategies when both strategies are sufficient for perfect performance.

Our previous simulations suggest that ALCOVE and RASHNL favor relatively independent predictors of the criterion. A form of independence that can arise with a very poor predictor of a criterion is the case of a constant predictor. A constant has a correlation of zero with the other predictors, and also with the criterion (very bad diagnosticity indeed). We explore whether ALCOVE and RASHNL have any attraction to this type or predictor.

There is reason to suspect that SUSTAIN may try to incorporate such a predictor. Although a constant dimension has zero correlation with other predictors, it will have maximal within-category consistency for any cluster. Thus, the inclusion of a constant dimension allows us to unconfound diagnosticity and between-predictor correlation from within-cluster consistency, possible aspects of the type of dimensions found to be attractive to SUSTAIN in previous simulations.

Inclusion of a constant dimension simulates important aspects of experimental stimuli that are usually ignored. The stimuli used in studies of category learning typically have many perceptually or conceptually salient aspects that are not coded or discussed by the experimenters, being treated as irrelevant because they are constant for all stimuli. For example, stimuli that are line drawings of bug-like creatures may differ in head shape, number of legs, and type of tail, aspects that are coded and manipulated by the experimenter to define the diagnostic input features to categorization models. But the line drawings all share certain basic characteristics that are constant across stimuli. Many models of similarity (e.g., Tversky, 1977; Markman & Gentner, 1993) assume that common features increase the similarity (and confusability) of stimuli. Thus, it seems interesting to use a simulation study to investigate what

predictions the three network models make for use of such constant or common-feature information.

Method: The category structure used for Simulation 2 is shown in Table 3. There are four exemplars of each category, A and B. Dimension D1 is a binary-valued variable, with values that are logically necessary-and-sufficient to identify each category. Dimension D2 is a constant dimension that has values of 1 for all exemplars in the population, regardless of category membership. Dimensions D3, D4, and D5 are binary-valued dimensions that together uniquely identify all eight exemplars. Note that this structure ensures that each network model not only has a relatively easy categorization strategy available (a unidimensional rule on D1), but can adopt a minimal attentional strategy that enables unique identification of all exemplars (attending to D3-D5).

Using the three models, we simulated subjects (N=100,000) who were trained for 20 blocks on the stimulus structure shown in Table 3. As in Simulation 1, for each individual subject parameter values were randomly selected from a uniform distribution within reasonable limits for each parameter. The main results recorded were the final-block attention weights for the five dimensions.

Table 3. Simulation 2: A simple two-category structure with one necessary-and-sufficient "category" dimension (D1), a constant dimension (D2), and three dimensions (D3-D5) that uniquely identify exemplars.

Class	D1	D2	D3	D4	D5
A	1	1	1	1	0
A	1	1	0	1	1
A	1	1	1	0	1
A	1	1	0	0	0
B	0	1	1	1	1
B	0	1	0	1	0
B	0	1	1	0	0
B	0	1	0	0	1

Results: Table 4 reports the mean pattern of relative attention in the final block for the successful classification learners, defined as those who had at least 80% classification accuracy in the final block.

Table 4. Mean final relative dimensional attention weights, by model, for the best-fitting simulated subjects of Simulation 2. Maximal mean attention weight for each model shown in bold type.

Model	D1	D2	D3	D4	D5
ALCOVE	**.338**	.097	.188	.188	.188
RASHNL	**.375**	.231	.132	.132	.132
SUSTAIN	**.389**	**.389**	.074	.075	.075

As can be seen in the table, learners simulated by ALCOVE gave the highest weight to D1, the dimension defining the simple rule. However, the total attention weight allocated by ALCOVE to the three dimensions uniquely identifying the exemplars (D3-D5) was greater than that given to the rule dimension D1, a pattern that could be interpreted as showing predominantly exemplar-based learning.[1] ALCOVE was relatively successful at ignoring the constant dimension D2, giving it about 1/4 the weight of the "rule" dimension D1. RASHNL showed a different pattern of final weights, giving the highest weight to the dimension (D1) defining the unidimensional category rule, an intermediate level to the constant dimension D2, and the least attention to the exemplar-identifying dimensions D3-D5. RASHNL's capability to emphasize D1, the rule dimension, is consistent with its capability to model simple rule-based strategies in other simulations we have conducted. It is somewhat surprising that this model cannot learn to ignore the constant dimension D2. SUSTAIN gave the least weight of any model to the "exemplar" dimensions D3-D5, and roughly as much weight as RASHNL to the perfectly diagnostic D1, but was the worst at ignoring D2, the constant dimension, giving it equal weight with D1.

Examination of the pattern of attention results across different regions of the parameter space for each model revealed that one key parameter affecting the results is the learning rate for association weights in the network. In order to display these results, we have created plots of the final pattern of attention weights for each model, separately for different ranges of the learning rate parameter.

Figure 1 presents the results for ALCOVE. The left panel plots the final attention weight for D2 (the constant dimension) versus that for D1 (the rule dimension). It can be seen that ALCOVE does not completely ignore D2 at any value of the learning rate, although D2 is consistently given lower weight than D1. The right panel plots the summed final attention weights for D3-D5, the "exemplar" dimensions, versus the weight for D1. These plots show a strong and consistent effect of the learning rate for associations. For higher values of this parameter (the upper row of the panel), the total attention weight given to the exemplar dimensions tends to exceed that for D1, meaning that exemplar learning predominates. For lower values of the learning rate (the bottom row) the dimension defining the unidimensional rule (D1) is weighted highly, sometimes even exclusively, meaning that a rule-based strategy is being used.

Figure 2 presents the corresponding plots for RASHNL. The left panel shows that RASHNL has trouble ignoring D2, the constant dimension, at any value of λ_w. However, D1 (the rule dimension) tends to receive more attention than D2 in the majority of solutions. The right panel shows that most simulated subjects pay more attention to D1, the rule dimension, than to the exemplar dimensions. This is especially true when the learning rate is very low (bottom row). However, the bottom row of the left panel

[1] Support for this interpretation is given by supplementary simulations we have conducted, in which various numbers of dimensions (1, 2 or 3) are used to uniquely code the exemplars. Across all of these simulations, the total final weight given to these "exemplar" dimensions is roughly constant, regardless of the number of dimensions involved.

underscores that for the low learning rates, considerable attention is also paid to D2, the constant dimension.

Figure 3 shows that SUSTAIN yields a very different pattern of results for this structure. For all values of λ_w SUSTAIN predicts that equal attention will be paid to D1 (the rule dimension) and D2 (the constant dimension). Also, the total amount of attention directed at D3-D5, the "exemplar" dimensions, is fairly stable across values of the learning rate, but there is more variability at the higher learning rates. Interestingly, the apparent constraint that the weights given to D1 and D2 be equal is so strong that any increase or decrease in the total weight given to D3-D5 trades off against the summed relative weight given to D1 and D2, creating a line of possible solutions with a slope of -2 in each plot of the right-hand panel.

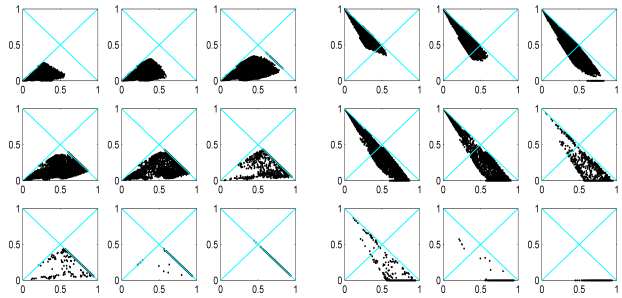

Figure 1. Simulation 2: Final relative attention weights for ALCOVE, separately for different values of λ_w, the learning rate for network association weights. Left panel: D2 (y-axis) versus D1 (x-axis) attention weights. Right panel: summed attention weights for D3, D4 & D5 (y-axis) versus D1 (x-axis) attention weights. In each panel, the nine plots summarize results for various ranges of the λ_w parameter. Top row: (>.8; .8-.4; .4-.2). Middle row: (.2-.15; .15-.10; .10-.05). Bottom row: (.05-.025; .025-.125; <.125).

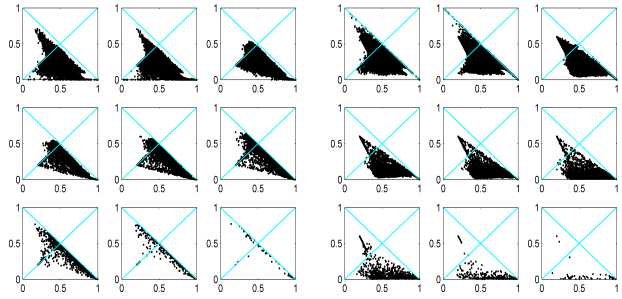

Figure 2. Simulation 2: Final relative attention weights for RASHNL

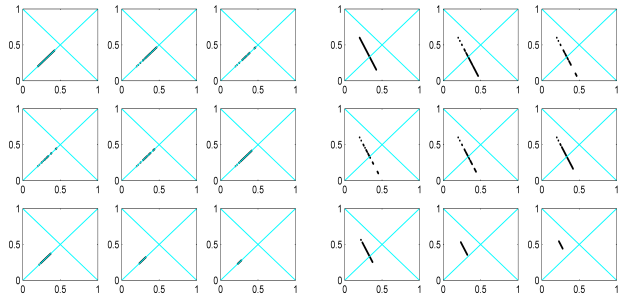

Figure 3. Simulation 2: Final relative attention weights for SUSTAIN

Discussion: The results of Simulation 2 are striking. First, both RASHNL and SUSTAIN pay considerable attention to a constant dimension (that has zero diagnosticity) under a wide range of parameter settings. In fact, RASHNL shows many solutions with relative weight exceeding 50% for D2 (with 5 dimensions). SUSTAIN invariably gives equal attention weight to D2 and D1, the unidimensional rule dimension. In this sense it is the least successful of the three models at ignoring D2. An explanation for this behavior of SUSTAIN is given below.

Second, the network models also differ in their tendencies to adopt the rule-based solution (using dimension D1) versus the exemplar-level representation (using D3-D5). For ALCOVE, successful learners tend to give high total attention weight to the "exemplar" dimensions D3-D5. These exemplar-based attention patterns occur often when the association learning rate is high, but rule-based attention patterns predominate when it is very low (Figure 1). For RASHNL, successful learners tend to weight the simple rule dimension (D1) more than the exemplar dimensions (D3-D5), and this tendency increases for low learning rates. Of the three models, SUSTAIN's successful learners give the least attention to the exemplar-identifying dimensions D3-D5. SUSTAIN pays somewhat more attention to these exemplar-identifying dimensions when the association learning rate is very low, the opposite pattern to that shown by ALCOVE and RASHNL.

Surprisingly, SUSTAIN gave the same amount of attention to D2 as to D1. Clearly, this tendency of SUSTAIN must arise from the structure and processing assumptions of the model. In fact, the reason that SUSTAIN finds D1 and D2 equally compelling is easy to identify, and stems from how SUSTAIN utilizes its reference points (i.e., clusters or prototypes) in learning. SUSTAIN utilizes only the single most activated cluster to determine an exemplar's classification and to guide learning. In this model, the update in attention strength for each dimension is inversely proportional to the distance from the most activated cluster's mean value and the value of the current input stimulus on that dimension (i.e., the smaller the dimensional distance, the more attention is increased for that dimension). For a constant dimension, any cluster and any input stimulus will have zero distance on that dimension, thus attention will be increased to the maximal degree possible on the constant dimension. In the present simulation, D1 is a perfect predictor with constant values *within* categories, thus any cluster that does not combine exemplars from across categories will also have zero distance on that dimension between the cluster centroid and the input stimulus, leading to an equivalent increase in attention strength to D2.

The critical aspect of the processing assumptions here is the winner-take-all nature of the utilization of the clusters, which means that the diagnosticity of a dimension relative to contrasting clusters has less effect. The net result in statistical terms is that the potential increase in attention to a dimension is a function of the similarity of the input

stimulus and the cluster centroid on that dimension. This places greater emphasis on within-category similarity and less on between-category differentiation, relative to the processing assumptions of ALCOVE and RASHNL. This line of analysis suggests that SUSTAIN will tend to select dimensions whose values have high category validity, P(f|c), over those with high cue validity, P(c|f), or with the best information gain (cf. Corter & Gluck, 1992).

Failing to ignore a dimension with zero diagnosticity seems like a major flaw of the three models, at least from a normative standpoint, because incorporating a constant dimension in a category's representation has cost without any obvious adaptive value. However, human data is needed to see if constant dimensions are indeed attended to and incorporated into a category's representation. It seems unlikely that in a category learning experiment human learners would waste time and effort memorizing or checking properties of a stimulus if those properties were seen to be useless for the task at hand.

On the other hand, it might be that such constant properties are learned implicitly, whether or not they are useful in a specific experimental task. An example might indicate why this is a reasonable possibility. A child learning the category *animal* might notice that all animals have mass. Is this fact incorporated into the child's representation? This certainly seems reasonable, though some normatively motivated theories of mental organization (e.g., Collins and Quillian, 1969) hold that the property of having mass should be stored at a superordinate level (say, under the category *object*) and merely inferred as needed in order to reason about animals and their properties.

Conclusions

The present analyses and simulation results show that the models examined here, ALCOVE, RASHNL, and SUSTAIN, incorporate differing attention learning mechanisms and processing assumptions that lead to distinct predictions regarding attention learning in the simulation studies reported here. The results from Simulation 1 supported the hypothesis that SUSTAIN tends to attend to dimensions that are correlated with other predictors, while the other models give relatively greater attention to more independent predictors, perhaps because they better support exemplar-level processing. Simulation 2 showed that the three models differ in their tendencies to use rule-based versus exemplar-based learning strategies. Another surprising result from Simulation 2 was that all three models incorporated a constant (i.e., completely nondiagnostic) dimension into their representations to some degree.

We believe that simulation studies on attention allocation in category learning are valuable for two reasons. First, they help us to better understand the behavior of complex computational models of category learning. Second, they can help to guide empirical work on attention by suggesting new hypotheses about human attention learning, hypotheses that can be verified using methods for assessing attention such as eye-tracking (e.g. Rehder & Hoffman, 2005) or information-board methods (Matsuka & Corter, 2008). These hypotheses may then be used to design empirical studies by suggesting stimulus structures and tasks that best differentiate predictions of the models.

References

Collins, A. M., & Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior, 8(2),* 240-247.

Corter, J. E., & Gluck, M. A. (1992). Explaining basic categories: Feature predictability and information. *Psychological Bulletin, 111(2)*, 291-303.

Corter, J.E., Matsuka, T., & Markman, A. B. (2007). Attention allocation in learning an XOR classification task. Poster presented at the Second European Cognitive Science Conference (*EuroCogSci 2007*), Delphi, Greece, May 23-27.

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

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

Love, B. C., & Medin, D. L. (1998). SUSTAIN: A model of human category learning. *Proceeding of the Fifteenth National Conference on AI (AAAI-98),* 671-676.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A network model of human category learning. *Psychological Review*, 111(2), 309-332.

Markman, A. B., & Gentner, D. (1993). Splitting the differences: A structural alignment view of similarity. *Journal of Memory and Language*, *32*, 517-535.

Matsuka, T., and Corter, J.E. (2008). Process tracing of attention allocation during category learning. *Quarterly Journal of Experimental Psychology, 61(7),* 1067-1097.

Matsuka, T., Corter, J. E., & Markman, A. B. (2002). Allocation of attention in neural network models of categorization. *Proceedings of the 24th Annual Meeting of the Cognitive Science Society*. Hillsdale, NJ: Lawrence Erlbaum.

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

Rehder, B., & Hoffman, A. B. (2005). Thirty-something categorization results explained: Selective attention, eyetracking, and models of category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31(5),* 811-829.

Tversky, A. (1977). Features of similarity. *Psychological Review, 84,* 327-352.