Psychological Review
2004, Vol. 111, No. 2, 309–332

0033-295X/04/$12.00 DOI: 10.1037/0033-295X.111.2.309

SUSTAIN: A Network Model of Category Learning

Bradley C. Love
University of Texas at Austin

Douglas L. Medin
Northwestern University

Todd M. Gureckis
University of Texas at Austin

SUSTAIN (Supervised and Unsupervised STratified Adaptive Incremental Network) is a model of how humans learn categories from examples. SUSTAIN initially assumes a simple category structure. If simple solutions prove inadequate and SUSTAIN is confronted with a surprising event (e.g., it is told that a bat is a mammal instead of a bird), SUSTAIN recruits an additional cluster to represent the surprising event. Newly recruited clusters are available to explain future events and can themselves evolve into prototypes–attractors–rules. SUSTAIN's discovery of category substructure is affected not only by the structure of the world but by the nature of the learning task and the learner's goals. SUSTAIN successfully extends category learning models to studies of inference learning, unsupervised learning, category construction, and contexts in which identification learning is faster than classification learning.

There is plenty of evidence to suggest that the key to the psychology of categorization is the flexible search for structure. Since Rosch's (e.g., Rosch, 1975; Rosch & Mervis, 1975) seminal studies of natural object categories, the scholarly consensus has been that, relative to our perceptual and conceptual systems, the world comes in natural chunks. That is to say, rather than comprising orthogonal distributions of features, the structure of things in the world consists of patterns of correlated features that create discontinuities or clusters (see also Berlin, Breedlove, & Raven, 1972). These clusters may provide the basis for cross-cultural agreement in categorization schemes (e.g., Malt, 1995) and tend to correspond to young children's assumptions about the extensions of category names (Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976). Even the view that categories are organized by theories requires that the theories be attuned to the affordances provided by the environment, if the theories are to be useful (Murphy & Medin, 1985).

But the search for structure must be flexible. First, even basic-level categories may have correlated features pointing to meaningful substructure. Second, people learn about and use hierarchically organized categories, so conceptual schemes must both coordinate and adjust to these different levels (see Waxman, 1998, for a review of developmental studies on this coordination task). Third, concepts and categories serve multiple functions, and the structure dictated by one goal may not be the most useful under some other goal or function (Solomon, Medin, & Lynch, 1999). Fourth, although our perceptual system has evolved, in part, to deliver useful categorizations, sometimes the categories suggested by perceptual similarity are far less useful than those that might be derived from a different analysis or weighting of features (e.g., Goldstone, Schyns, & Medin, 1997). Thus, the categorization system must be able to both assimilate structure and discover or even create that structure.

In this article, we introduce and describe experiments that explore a new model of category learning that is focused on the flexible search for structure: SUSTAIN (Love, Markman, & Yamauchi, 2000; Love & Medin, 1998a, 1998b). SUSTAIN (Supervised and Unsupervised STratified Adaptive Incremental Network) initially looks for simple solutions to category learning problems but is capable of entertaining more complex solutions when the problem calls for it. The category structures that SUSTAIN acquires are governed by both the structure of the world and the current task or goal.

The remainder of the article is organized as follows. First, we focus on category substructure and its implications for the power and flexibility of category learning models. Next, we describe SUSTAIN in terms of a series of general principles and present SUSTAIN's algorithm (i.e., the mathematical equations that follow from SUSTAIN's general principles). We then compare SUSTAIN with previous models of category learning. Next, we briefly overview the data sets SUSTAIN will fit, the majority of which are problematic for other models of category learning. In this analysis, we explain why SUSTAIN succeeds and why alternative models fail. Last, we summarize and consider the general

Bradley C. Love and Todd M. Gureckis, Department of Psychology, University of Texas at Austin; Douglas L. Medin, Department of Psychology, Northwestern University.

This work was supported by Air Force Office of Scientific Research Grant F49620-01-1-0295 to Bradley C. Love and National Institutes of Health Grant MH55079 and National Science Foundation Grant 9983260 to Douglas L. Medin. We thank F. Gregory Ashby, Jerome Busemeyer, John Kruschke, Levi Larkey, Todd Maddox, Art Markman, Greg Murphy, Thomas Palmeri, Paul Reber, Terry Regier, Lance Rips, Jeffrey Rouder, Yasu Sakamoto, Satoru Suzuki, and James Tanaka for their helpful comments.

Correspondence concerning this article should be addressed to Bradley C. Love, Department of Psychology, University of Texas at Austin, Austin, TX 78712. E-mail: love@psy.utexas.edu

implications of the SUSTAIN framework. The key contribution of SUSTAIN is to successfully extend models of category learning to a number of paradigms where other models either have not been applied or lead to incorrect predictions.

Flexibility and the Importance of Category Substructure

One challenge a human learner faces is uncovering the appropriate substructures within categories. Learning the substructure of a category enables the learner to both correctly classify instances of the concept and to make appropriate inferences. For example, even though both lions and horses are members of the category *mammals,* inferences that hold for lions may not hold for horses because these animals fall in different subcategories or conceptual clusters (felines vs. ungulates).

Learning the substructure of a category is not a trivial matter. The internal structure of a category can be highly nonlinear. For example, spoons tend to be large and wooden or small and made of steel. For the category *spoon,* there is not a characteristic weighting of the dimensions of material and size; rather, there are two distinct subgroups or conceptual clusters that contain opposite values on these two dimensions. Learning models that assume a simple category structure, such as prototype models (Posner & Keele, 1968), are unable to learn categories that have a rich internal structure. For example, the prototype for the category spoon would be situated (in representational space) between the large wooden spoons and the small steel spoons (Medin & Shoben, 1988). The prototype for the category spoon does not capture the distinct subtypes and would lead to inappropriate classifications and inferences. The prototype model is not an adequate model of human learning and category representation because it is too simple and inflexible.

In general, the complexity of the learner needs to be matched to the complexity of the learning problem. In the previous example, the complexity of the prototype model was insufficient to master the learning problem. Prototype models are biased only to learn categories that have a linear structure. Learning problems in which the decision boundary (in a multidimensional representational space) is highly irregular or in which there are multiple boundaries (e.g., all the members of a category do not fall inside one contiguous region of representational space) cannot be learned by a prototype model. Early neural network models (e.g., Rosenblatt, 1958) have similar limitations (Minsky & Papert, 1969).

More complex models can master nonlinear structures but may have difficulty with simpler structures. For example, a backpropagation model (Rumelhart, Hinton, & Williams, 1986) with many hidden units can learn complex decision boundaries but will perform poorly on a simple problem. For simple learning problems, overly complex models will tend to generalize poorly by overfitting the training data. Thus, making a model too powerful or too weak is undesirable. Geman, Bienenstock, and Doursat (1992) termed this tradeoff between data fitting and generalization as the *bias–variance dilemma.* In brief, when a model is too simple, it is overly biased and cannot learn the correct boundaries. Conversely, when a model is too powerful, it masters the training set, but the boundaries it learns may be somewhat arbitrary and highly influenced by the training sample, leading to poor generalization.

Flexible Power Through Incremental Adaptation

The complexity of learning models is usually fixed prior to learning. For example, in network models, the number of intermediate-level processing units (which governs model complexity) is usually chosen in advance (e.g., the number of hidden units in backpropagation model is set at the start of a simulation). The problem may not be avoidable by treating the number of intermediate units as an additional parameter, because certain architectures may be preferable at certain stages of the learning process. For example, Elman (1994) provided computational evidence (which seems in accord with findings from developmental psychology) that beginning with a simple network and adding complexity as learning progresses improve overall performance.

Ideally, a learner would adapt its complexity to the complexity of the learning problem. Indeed, some learning models have an adaptive architecture and adopt this approach. For example, some models begin large and reduce unneeded complexity (Busemeyer & McDaniel, 1997; Karnin, 1990), whereas other adaptive architecture models (including SUSTAIN) begin small and expand as needed (Ash, 1989; Azimi-Sadjadi, Sheedvash, & Trujillo, 1993; Carpenter, Grossberg, & Reynolds, 1991; Cho, 1997; Fahlman & Lebiere, 1990; Kruschke & Movellan, 1991).

Adaptive architecture learning models can be very effective in mastering a wide range of learning problems because they can adapt their complexity to the current problem. Humans face a similar challenge. Some categories have a very simple structure, whereas others can be complex. Accordingly, learning how to properly classify items as members of Category A or B can be almost trivial (e.g., when the value of a single input dimension determines membership) or can be so difficult that no regularity is discovered (e.g., rote memorization of every category member is required to determine membership). One possibility is that human learning follows the same trajectory, starting simple and adding complexity only as needed.

Multiple Goals and Functions

The analogy between machine learning and human learning can only be taken so far. The complexity of a machine learning problem can be equated with the complexity of the function that maps inputs (e.g., the stimulus to be classified) to outputs (e.g., the category membership of the stimulus). Human learning is not as easily (or as accurately) described in these terms alone.

For example, the category representation a human learner forms may be highly dependent on the current goals of the learner (e.g., Barsalou, 1985, 1991) and how categories are used (Love, 2003; Markman & Makin, 1998; Markman & Ross, 2003; Ross, 1996, 1997). Categories are often organized around these goals and conceptual structures are optimized to serve these goals (Medin, Lynch, Coley, & Atran, 1997). In a similar fashion, different conceptual functions (e.g., classification learning, inference learning, communication) all orient human learners toward different sources of information and may lead to different category representations, even when the structure of the information presented to the human learner is held constant. Depending on the task and learner's goals, the learner may spontaneously develop categories (so-called "unsupervised learning") or conceptual organization may be strongly constrained by feedback ("supervised learning").

A flexible model for learning about structure should be able to address a range of goals, tasks, and functions. As we show, SUSTAIN is able to do this.

The SUSTAIN model is intended as an account of how humans incrementally discover the substructure of categories. SUSTAIN matches its complexity to that of the learning problem but in a manner that is goal dependent and highly influenced by the learning mode engaged. These characteristics of SUSTAIN allow it to account for aspects of human learning that no other current model addresses.

Overview of SUSTAIN

SUSTAIN is a clustering model of human category learning. The basic components of the model are illustrated in Figure 1. Starting at the bottom of the figure, perceptual information is translated into a set of features that is organized along a set of dimensions. The example in the figure has values for shape, color, and the category label. Attentional tunings are learned for these dimensions. These tunings determine the importance of each feature dimension. The internal representations in the model consist of a set of clusters that are each associated with a category. The model attempts to assign a new instance to an existing cluster. This assignment can be done through an unsupervised learning procedure, although feedback can be used to determine if the initial assignment is correct. When the assignment is incorrect, a new cluster is formed to represent the current instance. Classification decisions are based on the cluster to which an instance is assigned.

Principles of SUSTAIN

SUSTAIN embodies five interrelated principles:

1. It is initially directed toward simple solutions.

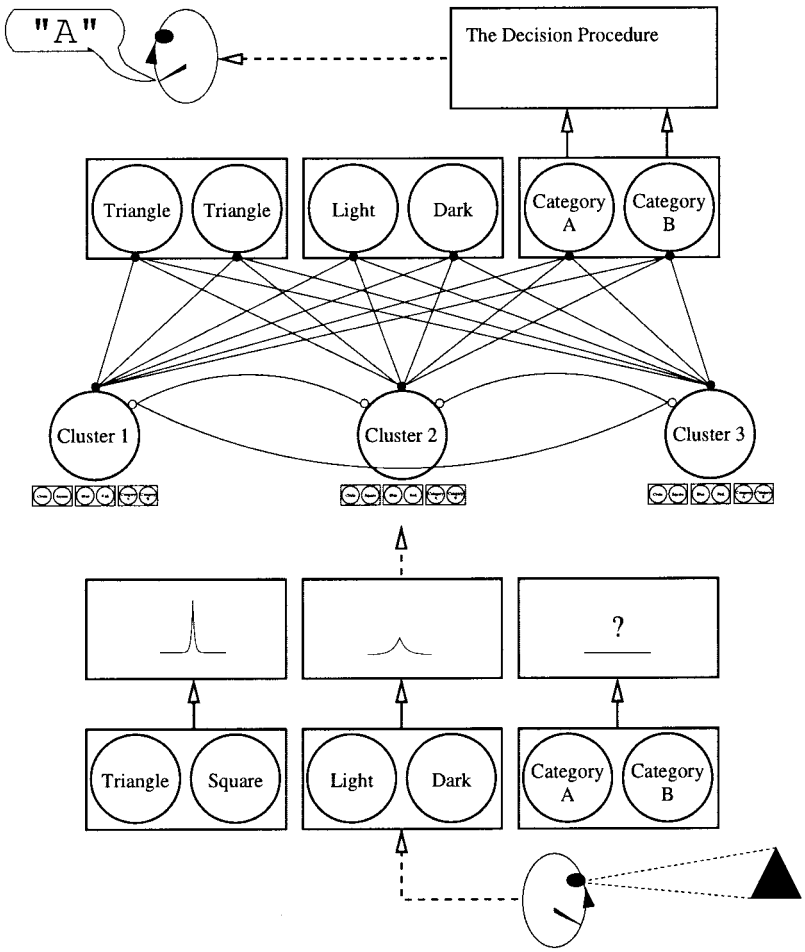

Figure 1. The basic components of the SUSTAIN model. First, the stimulus is encoded (in this case there are three binary-valued dimensions: two perceptual dimensions and the category label). The representational space is contorted (shrunk or stretched along each dimension) by the attentional mechanism. The clusters (in this case there are three) compete to respond to the stimulus. The cluster closest to the stimulus in representational space wins (through cluster competition; note the inhibitory connections among the three clusters). The winning cluster predicts the queried–unknown stimulus dimension value (in this case the category label) by sending a signal to the output units forming the queried dimension. These output units in turn serve as inputs to the decision procedure, which generates the response.

2. Similar stimulus items cluster together in memory.
3. Learning involves unsupervised and supervised processes.
4. Feedback affects the inferred category structure.
5. Cluster selection is competitive.

Principle 1: Simple First

SUSTAIN is initially directed toward simple solutions. SUSTAIN is biased toward simple solutions because it initially contains only one cluster and adds clusters (i.e., complexity) as needed. Its selective attention mechanism further serves to bias SUSTAIN toward simple solutions by focusing SUSTAIN on a subset of the possible stimulus dimensions that seems most predictive at the cluster level.

To illustrate SUSTAIN's preference for simple solutions, consider a classification learning problem in which animals must be segregated into Categories A and B. SUSTAIN would initially search for simple rules that segregate the stimuli into the two categories. For example, SUSTAIN would prefer solutions that involve one stimulus dimension (e.g., the items that *can fly* are in Category A, whereas the items that *cannot fly* are in Category B). When these simple solutions prove inadequate, more complex solutions involving multiple stimulus dimensions and exceptions are entertained.

There is one caveat—because SUSTAIN is an incremental clustering model, SUSTAIN can occasionally overlook a simple solution if the items are presented in an unfavorable order. Human learners are also susceptible to ordering effects (Bruner, Goodnow, & Austin, 1956; Garner & Whitman, 1965; Goldstone, 1996; Hovland & Weiss, 1953; Medin & Bettger, 1994). Ordering effects primarily arise from SUSTAIN's other principles (e.g., different item orderings lead to a different pattern of feedback, which affects the inferred category structure).

Principle 2: Similar Stimulus Items Tend to Cluster Together

In learning to classify stimuli as members of the category *birds* or *mammals,* SUSTAIN would cluster similar items together. For example, different instances of a bird subtype (e.g., sparrows) could cluster together and form a sparrow (or songbird) cluster instead of leaving separate traces in memory. Clustering is an unsupervised learning process because cluster assignment is done on the basis of similarity, not feedback. Although similarity drives clustering, clustering also drives similarity as attention shifts to stimulus dimensions that yield consistent matches across clusters.

Principle 3: SUSTAIN Is Capable of Both Supervised and Unsupervised Learning

In learning to classify birds and mammals, SUSTAIN would rely on both unsupervised and supervised learning processes. If SUSTAIN had a cluster whose members were small birds and another cluster whose members were four-legged mammals and SUSTAIN was asked to classify a bat, SUSTAIN would predict that a bat is a bird because the bat would be more similar to the small-bird cluster than to the four-legged mammal cluster (bats are small, have wings, fly, etc.). On receiving feedback (i.e., supervision) indicating that a bat is a mammal, SUSTAIN would recruit a new cluster to represent the bat stimulus.[1] In response to a prediction failure, SUSTAIN adds a cluster centered in representational space on the misclassified input. The next time SUSTAIN is exposed to the bat or another similar bat, SUSTAIN would correctly predict that a bat is a mammal. This example also illustrates how SUSTAIN can entertain more complex solutions when necessary (see Principle 1) through cluster recruitment.

An external oracle or teacher need not alert SUSTAIN to inappropriate clusterings. In cases in which there is no feedback (i.e., unsupervised learning), SUSTAIN is self-supervising. SUSTAIN recruits a new cluster (centered on the current example) when the similarity between the cluster most similar to the current item and the current item are below a threshold. In such cases, the most similar cluster does not strongly enough predict the current item and a new cluster is formed. This recruitment is analogous to the supervised process. Like the supervised case, SUSTAIN entertains complex solutions (involving numerous clusters) when necessary through cluster recruitment (driven by prediction or expectation failure). In both unsupervised and supervised learning situations, cluster recruitment is triggered by a surprising event.

Principle 4: The Pattern of Feedback Matters

As the example above illustrates, feedback affects the inferred category structure. Prediction failures on a queried dimension (e.g., the category label in classification learning) result in a cluster being recruited. Different patterns of feedback can lead to different representations being acquired. As we demonstrate later, this principle allows SUSTAIN to predict different acquisition patterns for different learning modes (e.g., inference vs. classification learning) that are informationally equivalent but differ in their pattern of feedback.

Principle 5: Cluster Competition

Clusters can be seen as competing explanations that attempt to explain the input. As such, the strength of the response of the winning cluster (the cluster the current stimulus is most similar to) is attenuated in the presence of other clusters that are somewhat similar to the current stimulus (cf. Sloman's, 1997, account of competing explanations in reasoning).

SUSTAIN's Formalization

The previous section presented the principles that underly SUSTAIN. These principles define SUSTAIN at an abstract level. This section explains how those general principles are manifested in an algorithmic model. The principles underlying SUSTAIN are more general than the equations that allow its predictions to be tested. The mapping from SUSTAIN's underlying principles to possible formalisms is likely many to one. The formalism presented here was chosen because it clearly reflects SUSTAIN's principles, allows predictions to be drawn readily, and facilitates comparisons with existing models. In the interests of these goals,

[1] Coincidentally, in some cultures bats are considered to be birds (see Lopez, Atran, Coley, Medin, & Smith, 1997).

SUSTAIN's formalism is idealized (i.e., simplified) when possible. The alternative path would yield a convoluted model containing numerous parameters and special conditions. This section is organized as follows: First, we specify SUSTAIN's input representation. Next, we discuss SUSTAIN's parameters. Finally, we present the equations that determine SUSTAIN's behavior.

Stimulus and Trial Representation

Stimuli are represented as vector frames where the dimensionality of the vector is equal to the dimensionality of the stimuli. The category label is also included as a stimulus dimension. Thus, stimuli that vary on three perceptual dimensions (e.g., size, shape, and color) and are members of one of two categories would require a vector frame with four dimensions. All simulations in this article involve nominal stimulus dimensions, as opposed to continuous stimulus dimensions (which SUSTAIN can also represent). A four-dimensional, binary-valued stimulus (e.g., three perceptual dimensions and the category label) can be thought of as a four-character string (e.g., 1 2 1 1) in which each character represents a stimulus dimension (e.g., the first character could denote the size dimension, with a *1* indicating a small stimulus and a *2* indicating a large stimulus). This notation is used throughout the article.

Of course, a learning trial usually involves an incomplete stimulus representation. For example, in classification learning all the perceptual dimensions are known, but the category-label dimension is unknown and queried. After the learner responds to the query, corrective feedback is provided. Assuming the fourth stimulus dimension is the category-label dimension, the classification trial for the above stimulus is represented as 1 2 1 ? → 1 2 1 1.

On every classification trial, the category-label dimension is queried and corrective feedback indicating the category membership of the stimulus is provided. In contrast, on inference learning trials, participants are given the category membership of the item but must infer an unknown stimulus dimension. Possible inference learning trials for the above stimulus description are as follows: ? 2 1 1 → 1 2 1 1, 1 ? 1 1 → 1 2 1 1, and 1 2 ? 1 → 1 2 1 1. Notice that inference and classification learning provide the learner with the same stimulus information after feedback (though the pattern of feedback varies).

Both classification and inference learning are supervised learning tasks. Unsupervised learning does not involve informative feedback. In unsupervised learning, every item is considered to be a member of the same category (i.e., the only category). Thus, the category-label dimension is unitary valued and uninformative.

To represent a nominal stimulus dimension that can display multiple values, SUSTAIN devotes multiple input units. To represent a nominal dimension containing k distinct values, k input units are used. All the units forming a dimension are set to zero, except for the one unit that denotes the nominal value of the dimension (this unit is set to one). For example, the stimulus dimension of marital status has three values (*single, married, divorced*). The pattern [0 1 0] represents the dimension value of married. A complete stimulus is represented by the vector $I^{\text{pos}_{ik}}$, where i indexes the stimulus dimension and k indexes the nominal values for Dimension i. For example, if marital status was the third stimulus dimension and the second value was present (i.e., married), then $I^{\text{pos}_{32}}$ would equal one, whereas $I^{\text{pos}_{31}}$ and $I^{\text{pos}_{33}}$ would equal zero. The pos in I^{pos} denotes that the current stimulus is located at a particular position in a multidimensional representational space. Note (see Figure 1) that SUSTAIN's output unit layer mirrors the input layer.

SUSTAIN's Parameters and Fit

SUSTAIN was simulated in a manner as consistent as possible with the procedures used in the original human experiments. For example, the same trial randomization procedures and stopping criteria were used for both human participants and SUSTAIN's simulations. Unlike the human results, which are averages of relatively small groups of individuals, SUSTAIN's performance is calculated by averaging over thousands of individual simulations to ensure that all means replicate to the level of precision reported.

Ideally, the variation observed across human participants would be compared with that across SUSTAIN simulations. Unfortunately, many of the original studies do not report human variance in a manner that allows for such comparisons to be made. In cases where these comparisons can be made, such as in the unsupervised studies considered here, SUSTAIN's predictions for variability are confirmed. Nevertheless, the focus of the fits is on group means.

For each human-study fit, the qualitative pattern of behavioral findings (supported by statistical tests of significance) is stated, as are the mean data (e.g., average number of blocks required to reach a learning criterion, overall accuracy, etc.). SUSTAIN's parameters are adjusted to minimize the sum of squared error among these data means and the means calculated by averaging over thousands of SUSTAIN simulations. A genetic algorithm is used to accomplish this minimization (Levine, 1996).

Although this parameter tuning improves SUSTAIN's fit, SUSTAIN's behavior is not extremely sensitive to the particular values of the parameters. There are certain behaviors that SUSTAIN cannot display no matter how the parameters are adjusted. The first 4 rows of Table 1 list SUSTAIN's basic parameters along with a brief description of the function of each parameter, the

Table 1
SUSTAIN's Best Fitting Parameters for All Data Sets Considered

Function/adjusts	Symbol	All studies	Six types	First/last name	Infer./class.	Unsupervised
Attentional focus	r	2.844642	9.01245	4.349951	1.016924	9.998779
Cluster competition	β	2.386305	1.252233	5.925613	3.97491	6.396300
Decision consistency	d	12.0	16.924073	15.19877	6.514972	1.977312
Learning rate	η	0.09361126	0.092327	0.0807908	0.1150532	0.096564
Category focus	λ_{label}	5.150151	—	—	12.80691	—
Distinct focus	$\lambda_{\text{distinct}}$	4.61733	—	5.213135	—	—

Note. Dashes indicate that the parameter was not applicable to the simulation. Infer. = inference; class. = classification.

symbol used to denote each parameter, and the value that provides the best fit for each study. These four parameters are used to fit the data in all studies.

Other parameters appear in particular studies. For example, in the unsupervised learning studies, SUSTAIN's cluster recruitment mechanism creates a new cluster when the current item is not sufficiently similar to any existing cluster. This threshold is captured by the parameter τ. The parameter τ can range between 0 and 1 but is somewhat arbitrarily set to .5 for all simulations. We chose the intermediate value to simplify the process of fitting and analyzing SUSTAIN. In other words, τ could be treated as a free parameter, but in the data fits presented here it is treated as a fixed value.

Some simulations demanded that the input presentation be parameterized because the original human learning study (from which the data simulated were drawn) did not equate the saliency of a feature dimension to that of the other dimensions. This stimulus parameter allowed SUSTAIN to alter the initial saliency of the uncontrolled dimension. In all cases, the values of these parameters appear sensible.

To demonstrate that SUSTAIN can simultaneously account for all of the key findings, a set of parameters (see Table 1 under the heading "All studies") was uncovered that allows SUSTAIN to capture the qualitative pattern of all of the studies reported in this article.[2] That SUSTAIN can capture the qualitative pattern of all of the studies with one set of parameters suggests that SUSTAIN's principles govern its performance rather than its specific parameter setting. It is important to note that the manner in which SUSTAIN fits each data set in the omnibus and individual fits is the same.

Mathematical Formulation of SUSTAIN

Each cluster has a receptive field for each stimulus dimension. A cluster's receptive field for a given dimension is centered at the cluster's position along that dimension. The position of a cluster within a dimension indicates the cluster's expectations for its members. The left panel of Figure 2 shows two receptive fields at different positions.

The tuning of a receptive field (as opposed to the position of a receptive field) determines how much attention is being devoted to the stimulus dimension. All the receptive fields for a stimulus dimension have the same tuning (i.e., attention is dimension-wide as opposed to cluster-specific). A receptive field's tuning changes as a result of learning. This change in receptive field tuning implements SUSTAIN's selective attention mechanism. Dimensions that are highly attended to develop peaked tunings, whereas dimensions that are not well attended to develop broad tunings. The right panel of Figure 2 shows two receptive fields with different tunings. Dimensions that provide consistent information at the cluster level receive greater attention.

Mathematically, receptive fields have an exponential shape, with a receptive field's response decreasing exponentially as distance from its center increases. The activation function for a dimension is

$$\alpha(\mu) = \lambda e^{-\lambda\mu}, \tag{1}$$

where λ is the tuning of the receptive field, μ is the distance of the stimulus from the center of the field, and $\alpha(\mu)$ denotes the response of the receptive field to a stimulus falling μ units from the center of the field. The choice of exponentially shaped receptive fields is motivated by Shepard's (1987) work on stimulus generalization.

Although receptive fields with different λs have different shapes (ranging from a broad to a peaked exponential), for any λ, the area underneath a receptive field is constant:

$$\int_0^\infty \alpha(\mu)d\mu = \int_0^\infty \lambda e^{-\lambda\mu}d\mu = 1. \tag{2}$$

For a given μ, the λ that maximizes $\alpha(\mu)$ can be computed from the derivative

$$\frac{\partial\alpha}{\partial\lambda} = e^{-\lambda\mu}(1 - \lambda\mu). \tag{3}$$

These properties of exponentials prove useful in formulating SUSTAIN.

With nominal stimulus dimensions, the distance μ_{ij} (from 0 to 1) between the *i*th dimension of the stimulus and Cluster *j*'s position along the *i*th dimension is

$$\mu_{ij} = \frac{1}{2}\sum_{k=1}^{v_i} |I^{\text{pos}_{ik}} - H_j^{\text{pos}_{ik}}|, \tag{4}$$

where v_i is the number of different nominal values on the *i*th dimension, *I* is the input representation (as described in a previous section), and $H_j^{\text{pos}_{ik}}$ is Cluster *j*'s position on the *i*th dimension for Value *k* (the sum of all *k*s for a dimension is 1). The position of a cluster in a nominal dimension is actually a probability distribution that can be interpreted as the probability of displaying a value given that an item is a member of the cluster. To return to a previous example involving marital status, a cluster in which 20% of the members are single, 45% are married, and 35% are divorced will converge to the location [.20 .45 .35] within the marital status dimension. The distance μ_{ij} will always be between 0 and 1 (inclusive).

The activation of a cluster is given by

$$H_j^{\text{act}} = \frac{\sum_{i=1}^m (\lambda_i)^r e^{-\lambda_i\mu_{ij}}}{\sum_{i=1}^m (\lambda_i)^r}, \tag{5}$$

where H_j^{act} is the activation of the *j*th cluster, *m* is the number of stimulus dimensions, λ_i is the tuning of the receptive field for the *i*th input dimension, and *r* is an attentional parameter (always nonnegative). When *r* is large, input units with tighter tunings (units that seem relevant) dominate the activation function. Dimensions that are highly attended to have larger λs and greater importance in determining the clusters' activation values. Increasing *r* simply accentuates this effect. If *r* is set to zero, every dimension receives equal attention. Equation 5 sums the responses of the receptive fields for each input dimension and normalizes the sum (again, highly attended dimensions weigh heavily). Cluster

[2] The decision-consistency parameter plays a minor role in the qualitative fit because its function is to scale the results (i.e., toward optimal behavior or chance guessing). Apart from extreme values that result in floor or ceiling effects, this parameter cannot alter SUSTAIN's qualitative predictions.

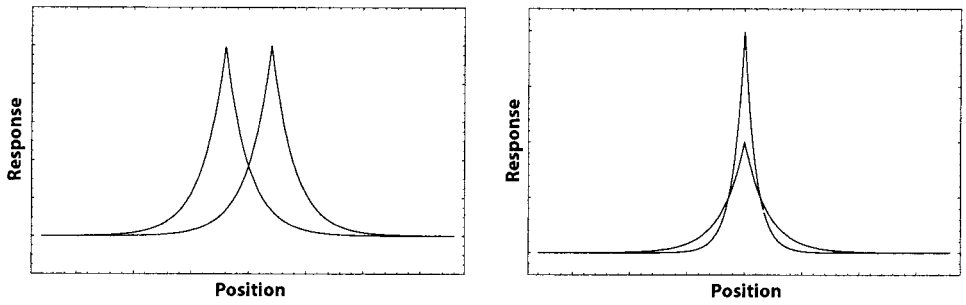

Figure 2. Left: Two receptive fields with different positions but the same tunings. Right: Two receptive fields with different tunings but the same positions. A maximal response is elicited when a stimulus falls in the center of a receptive field. Tightly tuned receptive fields (e.g., the taller receptive field in the right panel) can produce stronger responses, but their responses quickly diminish as distance from their center increases.

activation is bound between 0 (exclusive) and 1 (inclusive). Unknown stimulus dimensions (e.g., the category label in a classification trial) are not included in the above calculation.

Clusters compete to respond to input patterns and in turn inhibit one another. When many clusters are strongly activated, the output of the winning cluster H_j^{out} is less:

For the winning H_j with the greatest H^{act},

$$H_j^{\text{out}} = \frac{(H_j^{\text{act}})^{\beta}}{\sum_{i=1}^{n}(H_i^{\text{act}})^{\beta}} H_j^{\text{act}};$$

$$\text{for all other } H_j,\ H_j^{\text{out}} = 0, \quad (6)$$

where n is the number of clusters and β is the lateral inhibition parameter (always nonnegative) that regulates cluster competition. When β is large, the winner is only weakly inhibited. Clusters other than the winner are not selected and have their output set to zero. Equation 6 is a straightforward method for implementing lateral inhibition. It is a high-level description of an iterative process where units send signals to each other across inhibitory connections. Psychologically, Equation 6 signifies that competing alternatives reduce confidence in a choice (reflected in a lower output value).

Activation is spread from the clusters to the output units of the queried (the unknown) stimulus dimension *z:*

$$C_{zk}^{\text{out}} = \sum_{j=1}^{n} w_{j,zk} H_j^{\text{out}}, \quad (7)$$

where C_{zk}^{out} is the output of the output unit representing the kth nominal value of the queried (unknown) zth dimension, n is the number of clusters, and $w_{j,zk}$ is the weight from Cluster j to Category Unit C_{zk}. A winning cluster (especially one that does not have many competitors and is similar to the current input pattern) that has a large positive connection to a output unit will strongly activate the output unit. The summation in the above calculation is not really necessary given that only the winning cluster has a nonzero output but is included to make the similarities between SUSTAIN and other models more apparent.

The probability of making Response k (the kth nominal value) for the queried dimension z is

$$Pr(k) = \frac{e^{(d \cdot C_{zk}^{\text{out}})}}{\sum_{j=1}^{v_z} e^{(d \cdot C_{zj}^{\text{out}})}}, \quad (8)$$

where d is a response parameter (always nonnegative) and v_z is the number of nominal units (and hence output units) forming the queried dimension z. When d is high, accuracy is stressed and the output unit with the largest output is almost always chosen. The Luce (1959) choice rule is conceptually related to this decision rule.

After responding, feedback is provided to SUSTAIN. The target value for the kth category unit of the queried dimension z is

$$t_{zk} = \left\{ \begin{array}{l} \max(C_{zk}^{\text{out}}, 1), \text{ if } I^{\text{pos}_{zk}} \text{ equals } 1. \\ \min(C_{zk}^{\text{out}}, 0), \text{ if } I^{\text{pos}_{zk}} \text{ equals } 0. \end{array} \right\}. \quad (9)$$

Kruschke (1992) referred to this kind of teaching signal as a "humble teacher" (p. 2) and explained when its use is appropriate. Basically, the model is not penalized for predicting the correct response more strongly than is necessary.

A new cluster is recruited if the winning cluster predicts an incorrect response. In the case of a supervised learning situation, a cluster is recruited according to the following procedure:

For the queried dimension z,

if t_{zk} does not equal 1 for the C_{zk}

with the largest output C_{zk}^{out} of all C_{z*},

$$\text{then recruit a new cluster.} \quad (10)$$

In other words, the output unit representing the correct nominal value must be the most activated of all the output units forming the queried stimulus dimension. In the case of an unsupervised learning situation, SUSTAIN is self-supervising and recruits a cluster when the most activated cluster H_j's activation is below the threshold τ:

$$\text{If } (H_j^{\text{act}} < \tau), \text{ then recruit a new cluster.} \quad (11)$$

Unsupervised recruitment in SUSTAIN bears a strong resemblance to recruitment in adaptive resonance theory, Clapper and Bower's (1991) qualitative model (Carpenter & Grossberg, 1987), and Hartigan's (1975) leader algorithm.

When a new cluster is recruited (for both unsupervised and supervised learning situations), it is centered on the misclassified input pattern and the clusters' activations and outputs are recalculated. The new cluster then becomes the winner because it is the most highly activated cluster (it is centered on the current input pattern—all μ_{ij} will be zero). Again, SUSTAIN begins with a cluster centered on the first stimulus item.

The position of the winner is adjusted:

$$\text{For the winning } H_j,\ \Delta H_j^{\text{pos}_{ik}} = \eta(I^{\text{pos}_{ik}} - H_j^{\text{pos}_{ik}}), \quad (12)$$

where η is the learning rate. The centers of the winner's receptive fields move toward the input pattern according to the Kohonen (1982) learning rule. This learning rule centers the cluster amid its members.

Using our result from Equation 3, receptive field tunings are updated according to

$$\Delta\lambda_i = \eta e^{-\lambda_i\mu_{ij}}(1 - \lambda_i\mu_{ij}), \quad (13)$$

where j is the index of the winning cluster.

Only the winning cluster updates the value of λ_i. Equation 13 adjusts the peakedness of the receptive field for each input so that each input dimension can maximize its influence on the clusters. Initially, λ_i is set to be broadly tuned with a value of 1. The value of 1 is chosen because the maximal distance μ_{ij} is 1 and the optimal setting of λ_i for this case is 1 (i.e., Equation 13 equals zero). Under this scheme, λ_i cannot become less than 1 but can become more narrowly tuned.

When a cluster is recruited, weights from the unit to the output units are set to zero. The one-layer delta learning rule (Widrow & Hoff, 1960) is used to adjust these weights:

$$\Delta w_{j,zk} = \eta(t_{zk} - C_{zk}^{\text{out}})H_j^{\text{out}}, \quad (14)$$

where z is the queried dimension. Note that only the winning cluster will have its weights adjusted because it is the only cluster with a nonzero output. Equation 14 is somewhat idealized as it states that associations are only formed between the winning cluster and the output units of the queried dimension. In reality, some incidental learning to the output units of the nonqueried dimensions likely occurs. If of interest, such learning could be modeled by a second (lower) learning rate for nonqueried dimensions. For present purposes, we confine ourselves to the idealized version in the interest of avoiding a proliferation of parameters.

Comparing SUSTAIN With Other Category Learning Models

SUSTAIN is motivated by its own principles but nevertheless shares many commonalities with other models of category learning. Despite the commonalities, none of the models considered can account for the majority of the human learning studies that SUSTAIN is applied to later in this article.

The Configural Cue Model

Category learning in Gluck and Bower's (1988) configural cue model involves forming associations between a fixed feature set and output units. A category is defined by its associations with the input features. Associations are formed by an incremental learning process akin to linear regression (i.e., the one-layer delta learning rule). Unlike SUSTAIN, the configural cue model does not have intermediate units or an attentional mechanism. The input representation of the configural cue model consists of all possible combinations and subsets of combinations of all feature values (i.e., the power set). This mode of representation leads to computational problems. For example, to represent stimuli consisting of only three binary-valued feature dimensions (e.g., a large white triangle), the configural cue model needs 26 input units of which 7 are activated to encode a stimulus (e.g., large, white, triangle, large and white, large and triangle, white and triangle, large and white and triangle). The total number of input units required grows exponentially with the number of input dimensions, making the model untenable for problems with moderate dimensionality. For example, with 3 input dimensions that consist of binary features, the configural cue model needs 26 input units, but to represent 10 binary features the model needs 59,048 input units.

SUSTAIN input representation does not increase exponentially with the number of input dimensions because SUSTAIN discovers the relevant feature combinations and encodes them in its intermediate layer (i.e., as clusters). The combinations are discovered through unsupervised learning between the input and intermediate layer in conjunction with cluster recruitment. The selective attention mechanism plays a role in stressing which aspects of the clusters are most critical. Like the configural cue model, SUSTAIN uses the one-layer delta rule to adjust weights terminating at an output unit.

Rule-Plus-Exception (RULEX) Model

Although not apparent on the surface, there are deep commonalities between SUSTAIN and rule-based models like RULEX (Nosofsky, Palmeri, & McKinley, 1994). In trying to master a classification learning problem, RULEX first considers rules that are one dimensional (e.g., if Value 1 is present on the second dimension, then classify the item as a member of Category A). When a problem is not mastered by the rule, exceptions are encoded, or more complex rules are considered. Like RULEX, SUSTAIN first considers simple rules (i.e., solutions involving a small number of clusters), then encodes exceptions (i.e., additional clusters recruited through prediction failure), which can evolve into more complex rules. SUSTAIN's selective attention mechanism also biases it to initially search for simple rules that range over as few stimulus dimensions as possible. SUSTAIN's clusters can sometimes be interpreted as implementing rules (i.e., disjunctions of conjunctions in first-order logic). RULEX and SUSTAIN do differ in important ways, though. RULEX is a model of classification learning with two mutually exclusive categories. SUSTAIN is intended to be a more general model of learning. SUSTAIN's rule-like behavior is an emergent property that is displayed when mastering certain classification learning problems.

The Rational Model

Anderson's (1991) rational model is a clustering model. Like SUSTAIN, the rational model begins with one cluster and adds clusters incrementally. Both models attempt to capture and explicitly represent the substructure of categories. The rational model's principle goal is to uncover a cluster structure that captures statis-

tical regularities in the environment, whereas SUSTAIN recruits clusters in response to prediction failures. This distinction, although subtle, proves important. The rational model does not organize its knowledge structures around its current goals and task environment. The rational model's goal is always the same: to capture the statistical structure of the world. In addition to being sensitive to the structure of the world, SUSTAIN is also sensitive to the learning task and the current goals. For example, SUSTAIN can come up with two very different internal representations for a category depending on whether SUSTAIN is engaged in inference or classification learning. The rational model would not. A related point is that SUSTAIN, unlike the rational model, treats category labels or other dimensions that need to be predicted (i.e., that are queried) differently than nonqueried stimulus dimensions. For example, in classification learning, the category label plays an important role in directing the organizing of SUSTAIN's internal representations. Both the rational model and SUSTAIN seek to unify various learning modes under one model; a key difference is that SUSTAIN holds that different learning modes lead to different internal representations.

An important architectural difference between the two models is that the rational model is Bayesian. SUSTAIN makes predictions by focusing on the cluster that is most similar to the current item. The rational model makes predictions that are based on an optimally weighted sum over all clusters, instead of basing the response on the most active cluster. Recent work (Malt, Murphy, & Ross, 1995; Murphy & Ross, 1994) has suggested that SUSTAIN's focus on the most likely possibility may be in accord with human performance. Participants predict the value of a missing feature (loosely, this can be viewed as a category response) on the basis of the base-rate information of the most likely cluster, as opposed to a weighted sum across the probabilities of all clusters (as the rational model or any other optimal Bayesian approach does). SUSTAIN's ability to fit an array of data by only considering the most active cluster provides further support for the notion that humans may not fully consider alternative clusters after a winning cluster has been selected.

Abstract Approaches

Although the rational model is formulated at a fairly abstract level, it is nevertheless a model that contains parameters and learns on a trial-by-trial basis. Other models are even more abstract. For example, general recognition theory (Ashby & Townsend, 1986; Maddox & Ashby, 1993) does not attempt to characterize trial-by-trial learning but attempts to provide a concise description of human performance at asymptote. Such an approach can offer insights into human performance through comparison of human performance with that of an ideal observer (i.e., classifier). Although valuable, this approach (in the absence of auxiliary assumptions) does not provide an explanation for why human performance deviates from optimal or how learners reach asymptote.

Unlike SUSTAIN, the majority of abstract models are not formulated at the algorithmic level (i.e., many abstract models are not concerned with specifying the processes critical to human learning). Instead, many abstract models are computational level models (in the sense of Marr, 1982). These approaches view category learning as a function learning problem that maps inputs (i.e., stimuli) to outputs (i.e., categories). These approaches attempt to characterize the difficulty human learners will have with different functions or category partitions. Examples of computational level approaches include Corter and Gluck (1992), Feldman (2000), and Gosselin and Schyns (2001).

Computational level approaches are not suitable for addressing how different learning modes lead to different patterns of acquisition or the importance of goals in learning. These approaches are either not applicable to these questions or make incorrect predictions. SUSTAIN is formulated in way that allows it to address these questions. SUSTAIN is motivated by a set of abstract principles, but these principles are not solely concerned with the structure of the world. SUSTAIN is an important step in understanding how subjects (algorithmically) store and combine information about stimuli under a variety of learning conditions.

Exemplar-Based Approaches

Exemplar models (Hintzman, 1986; Medin & Schaffer, 1978; Nosofsky, 1986) store training instances in memory and classify stimulus items by computing the similarity of the current item to every previous exemplar. The item is then classified according to which exemplars it is most similar to overall (e.g., if a test item is very similar to many Category A members, an exemplar model will predict that the test item is a member of Category A). Exemplar models have been very successful as models of human category learning (see Estes, 1994). Exemplar models differ from SUSTAIN on a number of fronts. Perhaps the most important difference is that all abstraction in exemplar models is indirect, whereas in SUSTAIN it is direct. Exemplar models form abstractions by interpolating across the responses of many stored representations. In this regard, exemplar models are similar to the rational model (see Nosofsky, 1991a, for a comparison of the generalized context model, an exemplar model, and the rational model). In contrast, SUSTAIN focuses on the dominant cluster and directly stores its abstractions.

Exemplar models expand their internal representations with every training example, whereas SUSTAIN is more economical in its storage and only stores an example as a separate cluster when a prediction error occurs. Storage in SUSTAIN is therefore dependent on what is already stored in memory and the pattern of feedback (which allows SUSTAIN to predict that the same stimulus information can result in different internal representation when the learning mode or the current goals vary). The exemplar model most comparable with SUSTAIN is ALCOVE (Kruschke, 1992). ALCOVE blends connectionist learning rules with an exemplar category representation (i.e., the hidden units are exemplars). Like SUSTAIN, ALCOVE has a selective attention mechanism that orients it toward the most predictive stimulus dimensions. ALCOVE has been a very successful model of human learning. Because of its success and the fact that comparisons between ALCOVE and SUSTAIN serve to highlight SUSTAIN's properties, SUSTAIN's performance will be compared with ALCOVE's throughout this article. ALCOVE was fit to the data in the same manner as SUSTAIN. The Appendix provides details on ALCOVE for the interested reader. Best fitting parameter values are shown in Table A1 of the Appendix. Unlike SUSTAIN, ALCOVE was not fit to all studies simultaneously because it failed to account for all of the qualitative patterns of the studies in the individual fits.

Overview of the Human Data Sets Fit by SUSTAIN

This section provides a brief overview of the data sets to which SUSTAIN is applied. The majority of category learning research (and particularly research in category learning modeling) has exclusively focused on supervised classification learning. Category learning models have been able to fit data from this paradigm in impressive detail (e.g., Estes, 1994; Kruschke, 1992; Nosofsky, 1991b). We believe, however, that it is important for categorization models to address a range of tasks and conceptual functions. Although supervised category learning represents an important mode of acquisition to study, it is only one way out of many to learn about categories. Focusing exclusively on a single learning mode is a serious limitation for any theory that intends to explain category learning and generalization in any comprehensive sense (Love, 2001, 2002; Schank, Collins, & Hunter, 1986). Thus, the studies fit here, although making connections to foundational studies in classification learning, primarily focus on expanding the applicability of models to other induction tasks.

Collectively, the studies we review present a strong test of any model of human category learning. Only one of the following studies has been successfully fit by other models of category learning. Some of the studies involve learning procedures that are outside the boundary conditions of many category learning models. These studies address issues in category learning that are critical but have nevertheless not received a great deal of attention from modelers. SUSTAIN's fit of these data sets hinges on how its principles guide it toward uncovering category substructures (i.e., the clusters).

Foundational Classification Learning Findings

The first data set we consider is Nosofsky, Gluck, Palmeri, McKinley, and Glauthier's (1994) replication of Shepard, Hovland, and Jenkins's (1961) classification learning studies. In these studies, human participants learned to classify geometric stimuli into one of two categories (either Category A or B). Stimuli consisted of three perceptual binary-valued stimulus dimensions and a category label, which we view as the fourth stimulus dimension. The category label was queried on every trial and feedback was provided that indicated the correct category assignment. Six different mappings of stimuli to categories (i.e., six different learning problems) were examined and the challenge for models was to predict the relative difficulty of the different structures. Although some other learning models can fit Shepard et al.'s six classification learning problems, SUSTAIN's solution is novel and illustrates how SUSTAIN adapts its complexity to match the complexity of the learning problem. All other learning studies fit by SUSTAIN have proven difficult for other learning models to address.

Learning at Different Levels of Abstraction

The second study SUSTAIN fits is Medin, Dewey, and Murphy's (1983) studies comparing identification learning (learning in which each stimulus is assigned to its own singleton category) with category learning (many-to-one mapping of stimuli onto categories). Shepard et al. (1961) had also compared identification and categorization and found that identification learning appeared to represent an upper bound on the difficulty of categorization learning. Learning problems requiring item memorization should be more difficult than learning problems that promote abstraction, and many models of categorization are constrained to predict that categorization will be, at worst, no harder than identification. Unlike Shepard et al., Medin et al. used distinctive stimuli (photographs of faces) and found that identification learning was actually more efficient than classification learning. As we show, SUSTAIN offers an explanation for this counterintuitive finding. SUSTAIN's explanation is then tested in an experiment involving human participants that replicates and extends Medin et al.'s original study (Love, 2000).

Comparing Inference and Classification Learning

Inference learning is closely related to classification learning. In inference learning, the category label is known, but one of the perceptual dimensions is unknown and is queried. Like classification learning, inference learning is supervised and the learner receives corrective feedback. After receiving feedback the stimulus information available to the learner is equivalent in both inference and classification learning.

SUSTAIN was fit to a series of experiments (Yamauchi, Love, & Markman, 2002; Yamauchi & Markman, 1998) comparing human inference and classification learning. The basic finding was that inference learning promotes a focus on each category's prototype, whereas classification learning focuses human learners on information that discriminates between the categories. Accordingly, inference learning is more efficient than classification learning for linear category structures in which the category prototypes successfully segregate members of the contrasting categories but is less efficient than classification learning for nonlinear category structures in which the prototypes are of limited use. SUSTAIN is able to explain how these different patterns of behavior emerge from two learning tasks that are structurally equivalent.

Unsupervised Learning

In unsupervised learning, learners do not receive corrective feedback from an external oracle but are instead free to impose their own organization onto the stimulus set. In unsupervised learning, each stimulus may be viewed as belonging to the same category, and learners search for appropriate substructures to characterize the category. The idea is to see how learners spontaneously organize categories. SUSTAIN was fit to Billman and Knutson's (1996) human unsupervised learning studies. Billman and Knutson's studies explored how humans learn correlations among stimulus dimensions. In a series of experiments, Billman and Knutson found that intercorrelated structures (e.g., cor[A,B], cor[B,C]) were easier to learn than structures that were not intercorrelated (e.g., cor[A,B], cor[C,D]).[3] SUSTAIN prefers the category structures that human learners prefer.

SUSTAIN also addresses the unsupervised category construction studies of Medin, Wattenmaker, and Hampson (1987). In category construction (i.e., sorting studies), human participants are

[3] The term *cor(X,Y)* denotes that Dimensions X and Y have correlated values.

given cards depicting the stimuli and freely sort the cards into piles that naturally order the stimuli. In other words, human participants sort the stimuli into the natural substructures of the category. Medin et al. found (under several manipulations) that humans tend to create unidimensional sorts (e.g., place all the small stimuli in one pile and all the large stimuli in a second pile) even when the stimuli are intercorrelated across all stimulus dimensions and could be naturally partitioned into two piles that respect these intercorrelations. This finding serves as a counterpoint to Billman and Knutson's (1996) findings that demonstrate an advantage in learning intercorrelated structures. As we show, SUSTAIN reconciles this seemingly conflicting pattern of results.

Summary

The human data that SUSTAIN addresses in this article are drawn from classic studies in classification learning, studies in learning at different levels of abstraction, studies comparing classification and inference learning, and studies in unsupervised learning. The data considered cover a broad spectrum of category learning phenomena involving learning from examples. The total pattern of results may appear to give a fractured or even contradictory view of category learning. However, SUSTAIN provides a coherent view of these data that follows in a straightforward manner from its principles. The point of fitting SUSTAIN to these data sets is not to merely fit data that other models cannot fit but to increase our understanding of human category learning by highlighting the relations between these data sets.

Modeling Shepard, Hovland, and Jenkins (1961)

Shepard et al.'s (1961) classic experiments on human category learning provide challenging data to fit. Human participants learned to classify eight items that varied on three perceptual binary dimensions (shape, size, and color) into two categories (four items per category). On every trial, participants assigned a stimulus to a category and feedback was provided. Participants were trained for 32 blocks or until they completed 4 consecutive blocks without an error. For every study in this article, a block is defined as the presentation of each item in a random order. Six different assignments of items to categories were tested with the six problems varying in difficulty (Type I was the easiest to master, Type II the next easiest, followed by Types III–V, and Type VI was the hardest). This ordering of overall accuracy levels defines the qualitative pattern of results for Shepard et al. The logical structure of the six problems is shown in Table 2. The Type I problem only requires attention along one input dimension, whereas the Type II problem requires attention to two dimensions (Type II is XOR [exclusive-or] with an irrelevant dimension). The categories in the Type II problem have a highly nonlinear structure. Types III–V require attention along all three perceptual dimensions but some regularities exist (Types III–V can be classified as rule-plus-exception problems). Type IV is notable because it displays a linear category structure (i.e., Type IV is learnable by a prototype model). Type VI requires attention to all three perceptual dimensions and has no regularities across any pair of dimensions.

Nosofsky, Gluck, Palmeri, McKinley, and Glauthier (1994) replicated Shepard et al. (1961) with more human participants and traced out learning curves. Figure 3 shows the learning curves for the six problem types. The basic finding was that Type I is learned faster than Type II, which is learned faster than Types III–V, which are learned faster than Type VI. These data are particularly challenging for learning models because most models fail to predict that Type II is easier than Types III–V. The only models known to reasonably fit these data are ALCOVE (Kruschke, 1992) and RULEX (Nosofsky, Palmeri, & McKinley, 1994). SUSTAIN's fit of Nosofsky, Gluck, et al.'s (1994) data is shown in Figure 3. The procedure used to simulate SUSTAIN mimicked the procedure used to collect data from the human participants (i.e., random presentation of items in blocks, the same learning criterion, feedback on every trial, etc.). The best fitting parameters are shown in Table 1 under the heading "Six types."

Table 2
The Logical Structure of the Six Classification Problems Tested in Shepard et al. (1961)

	Classification category					
Stimulus	I	II	III	IV	V	VI
1 1 1	A	A	B	B	B	B
1 1 2	A	A	B	B	B	A
1 2 1	A	B	B	B	B	A
1 2 2	A	B	A	A	A	B
2 1 1	B	B	A	B	A	A
2 1 2	B	B	B	A	A	B
2 2 1	B	A	A	A	A	B
2 2 2	B	A	A	A	B	A

Note. The perceptual dimensions (e.g., large, dark, triangle, etc.) were randomly assigned to an input dimension for each participant.

How SUSTAIN Solves the Six Problems

SUSTAIN is not a black box and it is possible to understand how it solves a learning problem (perhaps providing insight into the problem itself). We now detail how SUSTAIN solves the six problems. The most common solution for the Type I problem is to recruit one cluster for each category. Type I has a simple category structure (the value of the first dimension determines membership). Accordingly, SUSTAIN solves the problem with only two clusters and shifts its attention almost exclusively to the first dimension (i.e., the value of λ for the first dimension is much larger than the value for the other two dimensions). Type II requires attention to two dimensions. SUSTAIN solves the Type II problem by allocating two clusters for each category. Each cluster responds to two input patterns, largely ignoring the irrelevant dimension. Because category members are highly dissimilar (e.g., 1 2 1 B and 2 1 2 B are in the same category), SUSTAIN forms two clusters for each category (ignoring differences on the irrelevant dimension).

Types III–V display a variety of solutions. The learning curves for Types III–V in Figure 3 reflect averaging over a family of solutions. Again, SUSTAIN is a trial-by-trial model of human category learning and incrementally uncovers the category structure of a classification problem. Different solutions arise (primarily) because different sequences of items occur on different training runs. For the Type III problem, the majority of solutions are of

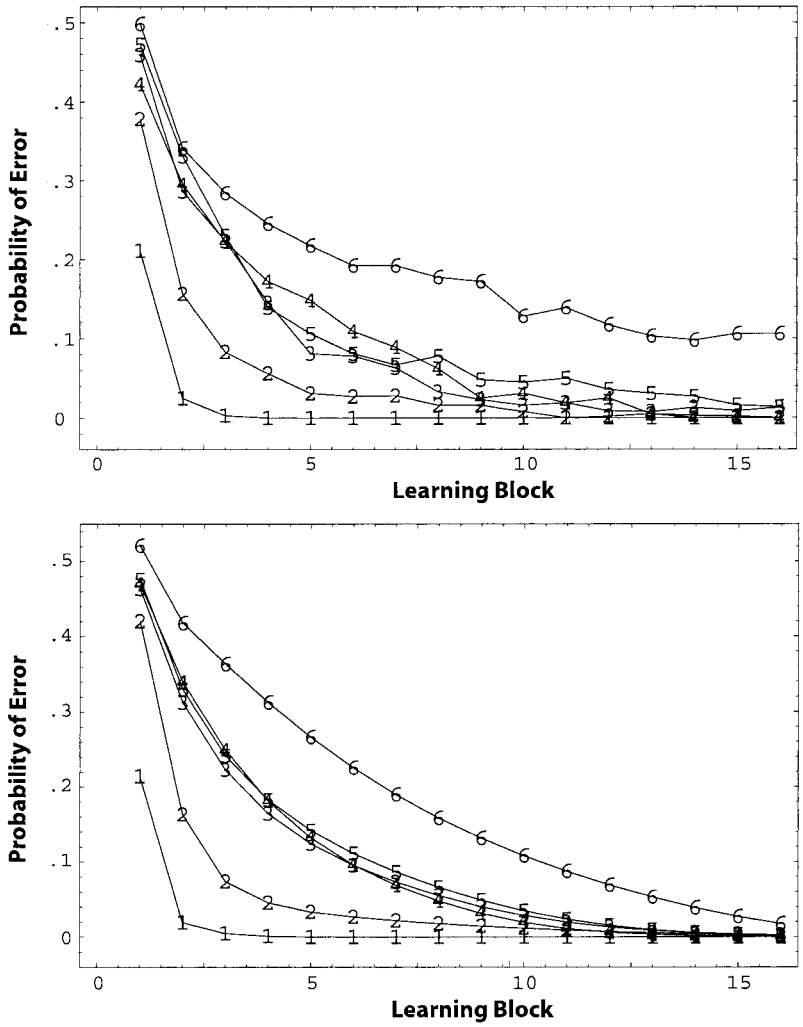

Figure 3. Top: Nosofsky, Gluck, et al.'s (1994) replication of Shepard et al.'s (1961) six problem types (labeled 1–6). Bottom: SUSTAIN's fit of Nosofsky, Gluck, et al.'s (1994) data (averaged over many simulations). The top panel is adapted from "Comparing Models of Rule Based Classification Learning: A Replication and Extension of Shepard, Hovland, and Jenkins (1961)," by R. M. Nosofsky, M. A. Gluck, T. J. Palmeri, S. C. McKinley, and P. Glauthier, 1994, *Memory & Cognition, 22,* p. 355. Copyright 1994 by the Psychonomic Society. Adapted with permission.

two varieties. The most common solution requires six clusters. Two clusters are created that each respond to two stimulus items (matching on the first 2 input dimensions). The remaining four clusters capture exceptions (i.e., each cluster is only strongly activated by one stimulus item). This solution allows attentional resources to be partially deployed to the first 2 dimensions. A less common solution only requires four clusters. Each cluster responds to two input patterns (matching on two dimensions). When this less common solution occurs, SUSTAIN masters the Type III problem more quickly than when the more common six-cluster solution arises.

SUSTAIN's most common solution for the Type IV problem is to recruit six clusters, with two of the clusters having two members each (again, clustered items have two input dimensions in common) and four clusters each encoding one stimulus item. The Type V problem is solved essentially the same way as the Type IV problem. One interesting difference between the Type IV and Type V problems is that SUSTAIN occasionally solves the Type IV problem with only two clusters (again, the modal solution to the Type IV problem requires six clusters). Runs displaying this infrequent solution reach learning criterion much faster than the modal Type IV solution. Although Type IV is a relatively difficult problem for people to master, the two-cluster solution is possible because a linear discriminant function (over all three perceptual dimensions) can separate the Category A and B items (i.e., any stimulus item in Table 2 with two or more input dimensions that have the first value is a member of Category B). Even when this rare two-cluster solution occurs because of a favorable ordering of training items, SUSTAIN still takes longer to master the Type IV problem than the Type I problem (the modal solution for the Type

I problem also uses two clusters) because SUSTAIN tends to prefer solutions that involve fewer dimensions. Humans also find unidimensional problems easier to master than other linear problems that require attending to multiple dimensions (Ashby, Alfonso-Reese, Turken, & Waldron, 1998; Ashby, Queller, & Berretty, 1999; Kruschke, 1993; McKinley & Nosofsky, 1996).

SUSTAIN solved the first 5 problem types by uncovering regularities and memorizing exceptions (devoting a unit for one item). Type VI has no regularities that can be exploited, forcing SUSTAIN to memorize each item (i.e., SUSTAIN devotes a cluster to each input pattern). In summary, for Shepard et al.'s (1961) six problems, the difficulty level of the problem is correlated with the number of clusters required to solve the problem. The modal solution to the Type I problem requires two clusters, Type II requires four clusters, Types II–V each require six clusters, and Type VI requires eight clusters. The Shepard et al. problems illustrate SUSTAIN's preference for simple solutions and how SUSTAIN matches its complexity to that of the learning problem. The fit also clearly illustrates how feedback affects the inferred category structure (all six problems involved the same eight items but with different patterns of feedback) and the interplay between unsupervised and supervised learning processes.

SUSTAIN's Principles: Item Versus Category Learning

As we have seen, the number of clusters SUSTAIN recruits varies with problem difficulty. For example, the most common solution for the Type I problem involves recruiting one cluster for each category. In contrast, the Type VI problem has no regularities that can be exploited, forcing SUSTAIN to memorize each stimulus (i.e., SUSTAIN devotes a cluster to each input pattern).

The Type VI problem is in some ways equivalent to identification learning (where each stimulus has a different label or category membership), whereas the Type I problem seems like a pure categorization problem (i.e., there is a simple criterion for membership; the categories are cohesive). It is tempting to conclude from the relative difficulty of the Type VI problem that identification learning is always more difficult than category learning. Contrary to this conclusion and to the predictions of other categorization models, there are striking instances where identification precedes categorization.

For example, Medin et al. (1983) found that people are faster to associate unique names to photographs of nine female faces than they are to categorize the photographs into two categories. The logical structure of the two categories is shown in Table 3 (the logical structure of the categories is roughly equivalent to Shepard et al.'s, 1961, Type IV problem). One key difference between the stimuli used in Medin et al.'s (1983) studies and in Shepard et al.'s studies that could have led to the identification learning advantage is that the stimuli used in Medin et al. were rich and distinct, varying along many dimensions not listed in Table 3, such as the shape of the face, the type of nose, and so forth. This *idiosyncratic* information makes each stimulus item more distinct.

SUSTAIN correctly predicts that the relative rates of identification and categorization learning interact with the nature of the stimuli. Specifically, when the stimuli are highly distinct, identification learning is faster than categorization. The properties of SUSTAIN that give rise to this behavior are discussed after simulation results are presented for Medin et al. (1983).

Table 3
The Logical Structure of the First-Name and Last-Name Conditions From Medin et al. (1983)

Stimulus	First name	Last name
1 1 1 2	A	A
1 2 1 2	B	A
1 2 1 1	C	A
1 1 2 1	D	A
2 1 1 1	E	A
1 1 2 2	F	B
2 1 1 2	G	B
2 2 2 1	H	B
2 2 2 2	I	B

Note. The four perceptual dimensions were hair color, smile type, hair length, and shirt color.

Modeling Medin et al. (1983)

Here, we focus on the first-name and last-name conditions from Medin et al. (1983). In the first-name condition participants learned a separate label for each photograph (i.e., identification learning), whereas in the last-name condition only two labels were used (category learning). The logical structure of the two conditions is shown in Table 3. In both conditions, participants were trained (using the classification learning procedure) until they correctly classified all nine items for consecutive blocks or until they completed the 16th learning block. Feedback was provided after each response.

The results from Medin et al. (1983) are shown in Table 4. The qualitative pattern of results is that more learning blocks (9.7 vs. 7.1) were required by participants in the last-name condition than in the first-name condition. Also, overall accuracy was roughly equal, even though chance guessing favored the last-name condition (i.e., pure guessing would result in 50% correct compared with 11% correct). When the first-name condition is rescored to account for guessing by scoring any label within the same category (A or B) as correct, overall accuracy rises to 91%, reliably higher than performance in the last-name condition.

To fit SUSTAIN to the data, certain assumptions had to be made about the nature of the input representation. Because participants were sensitive to the idiosyncratic information in each photograph, an additional input dimension was added to each item. The added dimension was nominal and displayed nine unique values (each stimulus displayed a unique value). Each stimulus's unique value on the added dimension represented the idiosyncratic information in each photograph (e.g., each person had a slightly different nose, shape of face, etc.). The added dimension had the effect of making each stimulus more distinctive. Of course, the saliency of this collective dimension was not matched to that of the four binary-valued perceptual dimensions in the original Medin et al. (1983) study. To account for the likely saliency differences, an additional parameter $\lambda_{\text{distinct}}$ was added to SUSTAIN. The additional parameter allowed SUSTAIN to initially weight the distinctive dimension differently than the other dimensions (dimensions normally have an initial λ of 1). In other words, we remained agnostic on the relative saliency of idiosyncratic information and allowed the model-fitting procedure to choose the desired level.

SUSTAIN was able to capture the correct pattern of results with this parameterization (see Table 4). The best fitting parameters are

Table 4
Human Performance and SUSTAIN's (in Parentheses) Scores for Medin et al. (1983)

Problem type	Blocks required	Overall accuracy
First name	7.1 (7.2)	.84 (.85)
Last name	9.7 (9.7)	.87 (.87)

shown in Table 1 under the heading "First/last name." It is worth noting that ALCOVE (with the added $\lambda_{distinct}$ parameter) could not predict a first-name advantage. Like people, SUSTAIN found it more natural to identify each stimulus than it did to associate several stimuli to a common label. SUSTAIN correctly predicts that overall accuracy between the two conditions should be roughly equal and that more learning blocks should be required in the last-name condition than in the first-name condition.

SUSTAIN recruited more clusters (nine for each simulation) in the first-name condition than in the last-name condition (the modal solution involved seven clusters). It is important to note that abstraction did not occur in the first-name condition (i.e., each cluster responded to only one item) but did occur in the last-name condition. When SUSTAIN's input representation did not include idiosyncratic information (i.e., the added stimulus dimension was removed), the last-name condition (blocks: 7.9; overall: .92) was easier to master than the first-name condition (blocks: 9.6; overall: .77). SUSTAIN predicts a strong interaction between stimulus distinctiveness and the learning task.

Why SUSTAIN Favors Identification Over Categorization in Medin et al. (1983)

Two factors conspire to cause SUSTAIN's performance to interact with the nature of the stimuli. As the stimuli become more distinctive, clusters that respond to multiple items are not as strongly activated. In other words, the benefit of abstraction is diminished with distinctive stimuli. This occurs because distinctive items sharing a cluster are not very similar to each other (i.e., within-cluster similarity is low). Notice that the diminished benefit of abstraction negatively impacts performance in the last-name condition but does not affect the first-name condition. In identification learning, each item forms its own cluster (within-cluster similarity is always maximal). When SUSTAIN is altered so that it does not form abstractions in either condition but instead recruits a cluster for each item, SUSTAIN fails to predict the interaction or the first-name condition advantage, suggesting that abstraction is critical for capturing this effect. Without abstraction, the inferred category structures (i.e., the clusters recruited) are identical for both conditions. Note that in exemplar models (which fail to capture the data), the internal representations for the first-name and last-name conditions are the same (nine exemplars), though the weightings of the exemplars differ.

The second factor that leads SUSTAIN to predict that distinctiveness and category level should interact is that the effects of cluster competition are attenuated with distinctive stimuli. As items become more distinctive, the clusters that are recruited tend to be further separated in representational space (i.e., the clusters match on fewer dimensions and mismatch on more dimensions). In other words, the clusters become more orthogonal to one another. The more distinctive that the clusters are, the less they tend to compete with one another. For example, when a distinctive stimulus is presented to SUSTAIN, it tends to strongly activate the appropriate cluster and only weakly activates the competing clusters. Reduced cluster competition with distinctive stimuli favors both identification and category learning but differentially benefits identification learning because there are generally more clusters present (i.e., potential competitors) in identification learning. Simulations support this analysis. When SUSTAIN is modified so that clusters do not compete, SUSTAIN reaches criterion more often and overall accuracy is higher in the last-name condition.

In summary, two factors, one related to abstraction and one to cluster competition, were responsible for SUSTAIN predicting that distinctiveness and category level should interact such that distinctiveness differentially favors identification learning over category learning. These results suggest that SUSTAIN may prove successful in explaining why certain categories are more natural or basic than others (Gosselin & Schyns, 2001; Rosch et al., 1976). For example, if asked how one gets to work in the morning, one says, "I drive my *car*," as opposed to "I drive my *Buick*," or "I drive my *vehicle*." SUSTAIN offers an explanation for why a level of categorization is preferred. In the above example, the intermediary category, *car*, balances the need to create clusters that have a high degree of within-cluster similarity and low degree of between-cluster similarity while minimizing the total number of clusters. Also, SUSTAIN's shift towards lower level categories in the presence of more distinctive inputs may be in accord with shifts in preferred category level with expertise (Johnson & Mervis, 1997; Tanaka & Taylor, 1991).

Further Tests of SUSTAIN's Account of Medin et al.'s (1983) Data

SUSTAIN's ability to fit Medin et al.'s (1983) studies on item and category learning is notable because other models cannot predict the advantage for identification learning or the interaction between learning task and stimulus distinctiveness. More important, SUSTAIN offers a framework for understanding the results. At the same time, it seems important to place SUSTAIN's account of these findings on firmer ground. To begin with, one should be cautious about accepting SUSTAIN's characterization of Medin et al.'s results. SUSTAIN's successful fit of Medin et al.'s studies depended on our choice of input representation. The idiosyncratic information in each photograph was represented by an additional input dimension. Each item had a unique value on the added dimension. This manipulation had the effect of making all the items less similar to one another.

The general intuition that guided our choice of input representation seems justified. Unlike artificial stimuli, the photographs vary along a number of dimensions. Still, replicating the results from Medin et al. (1983) under more controlled circumstances with artificial stimuli would bolster our claims. For example, it is possible that there may be something special about faces (cf. Farah, 1992), though there is evidence to the contrary suggesting that experience alone may be able to explain much of the data cited in favor of face-specific recognition systems (Diamond & Carey, 1986; Gauthier & Tarr, 1997; Rhodes, Tan, Brake, & Taylor, 1989). Nevertheless, humans do have a lot of experience in pro-

cessing faces and it is important to replicate the basic behavioral findings from Medin et al. with different stimuli.

Love (2000) conducted a study with human participants that directly supports SUSTAIN's account of Medin et al.'s (1983) data using artificial stimuli (schematic cars). Whereas Medin et al. featured a distinctive–identification learning condition (i.e., the first-name condition) and a distinctive–category learning condition (i.e., the last-name condition), Love included these two conditions along with a nondistinctive–identification learning condition and a nondistinctive–category learning condition, thus yielding a 2×2 factorial design: Learning Task (identification or category learning) $\times$ Stimulus Type (distinctive or nondistinctive). In the distinctive conditions, each stimulus was a unique color. In the nondistinctive conditions, each stimulus was the same color. SUSTAIN predicts that the learning task (identification or category learning) and the stimulus type (distinctive or nondistinctive) should interact such that identification learning will benefit more from distinctive stimuli than category learning. As in Medin et al., identification learning should be easier than category learning with distinctive stimuli. These predictions were confirmed.

Modeling Category Learning by Inference and Classification

Classification is clearly an important function of categories. Classifying an item allows category knowledge to be used. Inference is also a critical function of categories. For example, once we know a politician's party affiliation we can infer his or her stance on a number of issues. A number of studies have been directed at the way categories are used to make predictions (e.g., Heit & Rubinstein, 1994; Lassaline, 1996; Osherson, Smith, Wilkie, Lopez, & Shafir, 1990; Rips, 1975; Yamauchi & Markman, 2000).

In this section, we explore the different patterns of acquisition that result from classification and inference learning. As previously noted, the same information is available to the learner in classification and inference learning. The critical difference is that in inference learning, the learner is always given the category membership of a stimulus item and infers the value of an unknown perceptual dimension (the dimension queried varies across trials), whereas in classification learning the learner is always given the value of the perceptual dimensions and infers the category membership of the item. These two learning modes focus human learners on different sources of information and lead to different category representations. Inference learning tends to focus participants on the internal structure or prototype of each category, whereas classification learning tends to focus participants on information that discriminates between the two categories (Chin-Parker & Ross, 2002; Yamauchi et al., 2002; Yamauchi & Markman, 1998).

Accordingly, the difficulty of mastering a learning problem can be dependent on which of these two learning modes in engaged. The basic interaction observed between inference and classification learning is that inference learning is more efficient than classification learning for linear category structures in which the category prototypes successfully segregate members of the contrasting categories but is less efficient than classification learning for nonlinear category structures in which the prototypes are of limited use. Table 5 illustrates a linear category structure. Yamauchi and Markman (1998) trained participants on this category structure through either inference or classification learning. Learning was more efficient for inference learning than for classification learning. Conversely, when participants were trained on the nonlinear category structure from Yamauchi et al. (2002) shown in Table 6, classification learning was more efficient than inference learning. The complete pattern of results for these two studies is shown in Table 7. The qualitative pattern of results is the interaction between these two category structures and the induction task. In both studies, participants completed 30 blocks of training or until they surpassed 90% accuracy for a 3-block span.[4] SUSTAIN's runs were analyzed in the same fashion. The perceptual dimensions were form, size, color, and position.

The acquisition patterns for inference and classification learning for the linear and nonlinear category structure support the notion that inference learning focuses participants on the internal structure of each category, whereas classification learning focuses participants on information that discriminates between the categories. One interesting prediction that falls out of this characterization of inference and classification learning is that when the categories have a linear structure, inference learning should promote classification learning more than classification learning promotes inference learning. Uncovering the prototype of each category during inference learning provides a basis for subsequent classification learning (i.e., the key conceptual clusters have already been identified). Corroboration for this intuition can be found in SUSTAIN's simulations of the Type IV problem (which has a linear category structure). As we noted earlier, in the rare instances in which SUSTAIN only recruited one cluster for each category (i.e., the category prototype), learning was very efficient. In the majority of simulations, SUSTAIN learned the Type IV problem by focusing on discriminating information (i.e., creating imperfect rules and memorizing exceptions).

The reverse task ordering (classification learning followed by inference learning) should not be as efficient. If classification learning promotes a focus on discriminative rules and memorization of certain exemplars, inference learning should not benefit greatly from previous classification learning. The reason is that the representations acquired through classification learning are not appropriate for inference learning, which requires knowledge of the structure of the individuals categories (as opposed to the information that discriminates between the categories). Yamauchi and Markman (1998) tested and confirmed this prediction—inference learning followed by classification learning is the more efficient task ordering.

One important question is whether SUSTAIN can demonstrate the appropriate acquisition pattern for inference and classification learning. Yamauchi and Markman (1998) reported that the generalized context model (Nosofsky, 1986), which is an exemplar model like ALCOVE, and the rational model, which, like

[4] Blocks required to reach criterion were reported in a different fashion in the two studies. Participants who did not reach the learning criterion in Yamauchi et al.'s (2002) study were scored as "30," whereas such participants were excluded from analysis in Yamauchi and Markman (1998). In Yamauchi and Markman, 22 out of 24 participants reached criterion in inference learning and 23 out of 24 reached criterion in classification learning. The difference in data scoring did not prove critical because all SUSTAIN simulations for Yamauchi and Markman reached criterion.

Table 5
The Logical Structure of the Two Categories Tested in Yamauchi and Markman (1998)

Category A	Category B
1 1 1 0 A	0 0 0 1 B
1 1 0 1 A	0 0 1 0 B
1 0 1 1 A	0 1 0 0 B
0 1 1 1 A	1 0 0 0 B

SUSTAIN, forms clusters, have difficulty capturing the data from the linear category structures. A far greater challenge would be to account for the data from both the linear and nonlinear category structures.

Fitting SUSTAIN

The procedure used to train SUSTAIN mimicked the procedure used to train humans. The mean number of blocks required to reach criterion for each condition was fit (see Table 7). SUSTAIN's fit is also shown in this table. The best fitting parameters are shown in Table 1 under the heading "Infer./class." Note that an additional parameter, λ_{label} (category focus), was used in these simulations. The category focus parameter governs how much attention is placed on the category label at the beginning of a learning episode (akin to a participant's initial biases when entering the laboratory). Given the important organizational role that we hypothesize the category label plays (as well as the results from Yamauchi & Markman, 2000), we wanted to give SUSTAIN the option of placing more importance on the category label at the start of training. Following our intuitions, SUSTAIN differentially weighted the category label (see Table 1) relative to the perceptual dimensions, which had an initial tuning of 1.

Interpretation of the Model Fits

When engaged in inference learning with the linear category structure, SUSTAIN's modal solution was to recruit two clusters (one for each category). These two clusters were the prototypes of the two categories. Attention was greater (both initially and in end state) for the category-label dimension than for the perceptual dimensions. When engaged in classification learning with the linear category structure, SUSTAIN typically recruited six clusters to classify the eight items (i.e., SUSTAIN discovered some regularity and memorized a number of exceptions). SUSTAIN had a very difficult time when engaged in inference learning with the nonlinear category structure. In this case, SUSTAIN's focus on the category-label dimension was detrimental because the prototypes of each category were not sufficient to segregate the category members correctly. SUSTAIN's focus on the category label led to it recruiting 10 clusters, which is more clusters than there are items. In the case of classification learning, no salient regularity existed and SUSTAIN simply memorized the six items. SUSTAIN's modal solution is consistent with an account of human inference and classification learning that posits that inference promotes a focus on the internal structure of each category, whereas classification learning orients learners toward discriminative information.

Table 6
The Logical Structure of the Two Categories Tested in Yamauchi et al. (2002)

Category A	Category B
1 1 1 1 A	1 1 0 1 B
1 1 0 0 A	0 1 1 0 B
0 0 1 1 A	1 0 0 0 B

Table 7
The Mean Number of Inference and Classification Learning Blocks Required for Humans and SUSTAIN (in Parentheses)

Category structure	Inference	Classification
Linear	6.5 (7.5)	12.3 (11.2)
Nonlinear	27.4 (28.6)	10.4 (10.6)

The way SUSTAIN fits the inference and classification data also allows it to correctly predict that classification following inference is more efficient than the reverse ordering (for the linear category structure). When SUSTAIN displays the modal solution and recruits two clusters for inference learning, these two clusters are usually sufficient for successful classification learning. In other words, SUSTAIN can recycle its previous knowledge structures and simply learn to associate a new response with each cluster.

ALCOVE was also fit to the acquisition data. Somewhat to our surprise, ALCOVE successfully captured the pattern of data shown in Table 7. ALCOVE fit the data in Table 7 by placing a high initial weight on the category label (i.e., the λ_{label} had a high value). ALCOVE basically behaved like a prototype model. In inference learning on the linear category structure, ALCOVE's focus on the category label became very extreme in the end state. ALCOVE placed all of its attention on the category label and no attention on the perceptual dimensions. This allowed ALCOVE to implement a prototype model. Essentially, each stimulus strongly activates only the members of its category (which can be thought of as forming a distributed prototype cluster) and none of the items from the other category. Unfortunately, ALCOVE's lack of attention to perceptual dimensions is problematic for transfer to classification learning from inference learning. ALCOVE's weighting account of the tasks differs from SUSTAIN's, which posits that inference and classification learning lead to different category representations.

In the nonlinear case, ALCOVE's initial focus on the category-label dimension was detrimental because behaving like a prototype model is not advantageous when the prototypes do not sufficiently separate the category members. In this case, ALCOVE did not shift all of its attention away from the perceptual dimensions. In classification learning (with both category structures), ALCOVE focused on the discriminative perceptual dimensions that aided in correctly predicting the category label.

Explanatory Value, Falsifiability, and Model Complexity

Further simulations were conducted to determine whether SUSTAIN's and ALCOVE's fit of the data in Table 7 were explanatory rather than merely descriptive. If a model was suffi-

ciently complex that it could fit any possible pattern of results, it would not be impressive when the model fit the pattern of results displayed by humans.

To test this possibility, we generated a fictitious data set where classification learning was more efficient than inference learning for the linear category structure but less efficient for the nonlinear category structure. Note that this fictitious pattern of results is the opposite of what was observed. We then looked to see whether ALCOVE and SUSTAIN could fit these data; they could not. The inability of the models to account for a pattern of results that humans do not generate suggests that, although the models are highly parameterized, they are potentially falsifiable. Additional evidence showing that SUSTAIN is explanatory is that SUSTAIN correctly predicted the task ordering result despite only being fit to the data from the first learning task that participants completed and not the second.

Modeling Unsupervised Learning

In SUSTAIN there is no principled difference between supervised and unsupervised learning. In either case a cluster is recruited when a surprising event occurs. For supervised learning, the surprising event is an error (an incorrect prediction at the output layer). In unsupervised learning, errors cannot be made because there is no discriminative feedback (and each item is modeled as being a member of the same category). In unsupervised learning, the surprising event is a new stimulus that is not sufficiently similar to any stored representation (i.e., cluster). These two notions of surprise are quite compatible. In fact, a modified version of SUSTAIN uses a common recruitment mechanism for both unsupervised and supervised learning (Gureckis & Love, 2003).

Although unsupervised learning has not been as extensively studied as supervised learning, people can learn without external feedback (Homa & Cultice, 1984). One important challenge is to characterize how humans build internal representations in the absence of explicit feedback. To evaluate SUSTAIN's promise as a model of unsupervised learning, SUSTAIN was fit to a series of unsupervised learning studies. First, two studies from Billman and Knutson (1996) that explored the nature of unsupervised correlation learning were fit. Then, SUSTAIN was applied to unsupervised category construction (i.e., sorting) data from Medin et al. (1987).

Billman and Knutson's (1996) Experiments 2 and 3

Billman and Knutson's (1996) experiments tested the prediction that category learning is easier when stimulus dimensions are predictive of other dimensions (e.g., "has wings," "can fly," and "has feathers" are all intercorrelated). Broadly, their studies evaluated how relations among stimulus dimensions affect unsupervised learning.

Experiment 2 consisted of two conditions: nonintercorrelated and intercorrelated. In the nonintercorrelated condition, there was only one pairwise correlation between the perceptual stimulus dimensions, whereas in the intercorrelated condition there were six pairwise correlations. In the intercorrelated condition, the correlations were also interrelated (e.g., cor[A,B], cor[B,C], cor[A,C]). Stimulus items depicted imaginary animals consisting of seven perceptual dimensions: type of head, body, texture, tail, legs, habitat, and time of day pictured. Each dimension could take on one of three values (e.g., "sunrise," "midday," "nighttime"). In both conditions, participants studied the stimulus items (they were told that they were participating in an experiment on visual memory) for four blocks. This segment of the experiment served as the study or learning phase.

In the test phase of the experiment, participants viewed a novel set of 45 stimulus item pairs. Each member of the pair had two unknown (i.e., obscured) dimension values (e.g., the locations where the tail and head should have been were blacked out). Participants evaluated the remaining five perceptual dimensions and chose the stimulus item in the pair that seemed most similar to the items studied in the learning phase (a forced-choice procedure). One of the test items was considered the correct test item because it preserved one of the correlations present in the items viewed during the study phase. Table 8 shows the logical structure of the study and test items. The basic result from Experiment 2 was that

Table 8
The Logical Structure of the Studied Stimulus Items for the Nonintercorrelated and Intercorrelated Conditions in Experiments 2 and 3 of Billman and Knutson (1996)

Experiment	Nonintercorrelated	Intercorrelated
2	11xxxxx 22xxxxx 33xxxxx	1111xxx 2222xxx 3333xxx
3	111111x 221111x 331111x 111122x 221122x 331122x 111133x 221133x 331133x 112211x 222211x 332211x 112222x 222222x 332222x 112233x 222233x 332233x 113311x 223311x 333311x 113322x 223322x 333322x 113333x 223333x 333333x	111xxxx 222xxxx 333xxxx

Note. The seven columns within each cell denote the seven stimulus dimensions. Each dimension can display one of three different values, indicated by a 1, 2, or 3. An x indicates that the dimension was free to assume any of the three possible values.

the correct item was chosen more often in the intercorrelated condition than in the nonintercorrelated condition (73% vs. 62%).

Experiment 3 replicated Experiment 2's result and ruled out the possibility that the advantage of the intercorrelated condition in Experiment 2 was simply due to the greater number of pairwise correlations in the intercorrelated condition. The logical structure of the study and test phase is shown in Table 9. As in Experiment 2, the correct item was chosen more often in the intercorrelated condition than in the nonintercorrelated condition (77% vs. 66%). The advantage of the intercorrelated conditions over the nonintercorrelated conditions in these two experiments is the qualitative pattern of results.

Fitting SUSTAIN to Billman and Knutson's (1996) Experiments 2 and 3

In the supervised learning studies modeled in this article, participants (and SUSTAIN's) performance was measured in terms of accuracy or the number of learning blocks required to meet a criterion. In Billman and Knutson's (1996) studies, a participant's task was to observe a series of stimulus items without any feedback (the learning phase) and then (in the test phase) the participant made a series of decisions that involved choosing the more familiar stimulus item from a pair of stimulus items (a forced choice). SUSTAIN's task was to mimic the preferences participants displayed.

Equation 8 was used to model the forced-choice decisions. In deciding which of two test stimuli was most similar to previously studied items, the output of the output unit representing the category-label dimension (again, all items are assumed to be members of the same category) was calculated for both stimuli and these two values were used to calculate the response probabilities (i.e., in Equation 8 v_z now represented the number of alternatives in the forced choice and C_{zk}^{out} represented the output of the category unit in response to the kth item). During the test phase, unknown stimulus dimensions were modeled by setting the λ associated with that dimension to zero for the duration of the trial (i.e., unknown dimensions did not affect cluster activation).

SUSTAIN was trained in a manner analogous to how participants were trained. No feedback was provided and all stimulus items were encoded as being members of the same category. New clusters were recruited according to Equation 11. The best fitting parameters for both Experiments 2 and 3 (one set of parameters was used to model both studies) are shown in Table 1 under the heading "Unsupervised." SUSTAIN's fit is shown in Table 9.

SUSTAIN correctly predicted the preference ordering in both experiments. SUSTAIN, like humans, preferred intercorrelated stimulus dimensions and displayed greater accuracy for the intercorrelated than for the nonintercorrelated conditions. In Experiment 2, the stimuli in the intercorrelated condition were naturally partitioned into three groups defined by the correlated dimensions, which are ternary valued. Accordingly, SUSTAIN recruited three clusters and shifted its attention to the correlated stimulus dimensions that the clusters were organized around. In Experiment 2's nonintercorrelated condition, SUSTAIN's modal solution again involved three clusters organized around the correlated dimensions. However, the clusters in the nonintercorrelated condition were not as encompassing as in the intercorrelated condition. The clusters in the nonintercorrelated condition were organized around one pairwise correlation, whereas the clusters in the intercorrelated condition were organized around four intercorrelated dimensions, leading to higher accuracy levels in the intercorrelated condition.

Table 9
The Mean Accuracy for Humans and SUSTAIN (in Parentheses) for Billman and Knutson's (1996) Experiments 2 and 3

Experiment	Nonintercorrelated	Intercorrelated
2	0.62 (0.66)	0.73 (0.78)
3	0.66 (0.60)	0.77 (0.78)

The way SUSTAIN fit Experiment 3 parallels Experiment 2 with one interesting exception. Like Experiment 2, SUSTAIN's most common solution in the nonintercorrelated condition was to partition the studied items into three groups. Unlike Experiment 2, the nature of the three partitions varied across runs. SUSTAIN tended to focus on one of three correlations present in the nonintercorrelated condition and ignored the other two (a blocking effect was displayed). For example, during training SUSTAIN might create three clusters organized around the first two input dimensions (one cluster for each correlated value across the two dimensions) and largely ignore the correlation between the third and fourth dimensions and the fifth and sixth dimensions. The fact that the correlations are not interrelated makes it impossible for SUSTAIN to capture more than one correlation within a single cluster. SUSTAIN could recruit more clusters to represent all of the pairwise correlations, but instead SUSTAIN's bias toward simple solutions directs it to two of the seven dimensions (i.e., one of the pairwise correlations).

The same dynamics that led SUSTAIN to focus on only one correlation in the nonintercorrelated condition led SUSTAIN to focus on all of the interrelated correlations in Experiment 3's intercorrelated condition. When SUSTAIN learned one correlation in the intercorrelated conditions, SUSTAIN necessarily learned all of the pairwise correlations because of the way clusters were updated. SUSTAIN's solution suggests some novel predictions: (a) Learning about a correlation is more likely to make learning about another correlation more difficult when the correlations are not interrelated. (b) When correlations are interrelated, either all of the correlations are learned or none of the correlations are learned. These predictions have been verified (Gureckis & Love, 2002).

Other Models' Performance

SUSTAIN can capture the qualitative patterns in Billman and Knutson's (1996) data. Many other models cannot. For example, Billman and Knutson reported that certain exemplar models (e.g., Medin & Schaffer, 1978) and models that repeatedly probe instance memory (e.g., Heit, 1992; Hintzman, 1986) have problems capturing the qualitative pattern of results.

ALCOVE's fits did not converge on a unique solution, perhaps suggesting that ALCOVE's predictions are highly dependent on its specific parameter settings. Like SUSTAIN, one class of solutions showed the correct pattern of results. An example parameter set for such a solution is shown in Table A1 of the Appendix. Another popular solution was to favor category structures that contained fewer correlated dimensions. These solutions correctly predict higher accuracy for the intercorrelated condition in Experiment 3

but incorrectly predict higher accuracy for the nonintercorrelated condition in Experiment 2. A third common solution was to predict near equal performance in both conditions for both experiments. Given the large number of training items (each of which corresponds to a hidden unit in ALCOVE), it is difficult to understand ALCOVE's account of the data. In contrast, SUSTAIN's clustering account is interpretable and has proven useful in generating behavioral predictions.

Modeling Category Construction

In category construction (i.e., sorting studies), human participants are given cards depicting the stimuli and freely sort the cards into piles that naturally order the stimuli. In other words, participants sort the stimuli into the natural substructures of the category without any supervision. In Billman and Knutson's (1996) studies, we saw that participants preferred stimulus organizations in which the perceptual dimensions were intercorrelated. Category construction studies reveal a contrasting pattern: Participants tend to sort stimuli along a single dimension. This behavior persists even when alternate organizations respect the intercorrelated nature of the stimuli (Medin et al., 1987).

For example, Medin et al. (1987) found that participants tended to sort the stimulus set depicted in Table 10 along one of the four perceptual dimensions (e.g., participants placed all the stimuli with angular heads in one pile and all the stimuli with round heads in a second pile) even though there was a natural grouping of the stimuli that captured the intercorrelated family resemblance structure of the stimulus set (i.e., the stimuli in the left column of Table 10 in one pile and the stimuli in the right column in the second pile).

Modeling Sorting Behavior With SUSTAIN

SUSTAIN was applied to the sorting data from Medin et al.'s (1987) Experiment 1 in hopes of reconciling the apparent contradictory findings from category construction studies and Billman and Knutson's (1996) studies. In Experiment 1, participants were instructed to sort the stimuli into two equally sized piles. Stimuli were cartoon-like animals that varied on four binary-valued perceptual dimensions (head shape, number of legs, body markings, and tail length). The logical structure of the items is shown in Table 10. The basic finding was that participants sorted along a single dimension as opposed to sorting stimuli according to their intercorrelated structure (i.e., family resemblance structure).

SUSTAIN was applied to the stimulus set from Experiment 1. SUSTAIN, like Medin et al.'s (1987) participants, was constrained to only create two piles (i.e., clusters). This was accomplished by not allowing SUSTAIN to recruit a third cluster.[5] SUSTAIN was presented with the items from Table 10 for 10 training blocks. The multiple blocks were intended to mirror human participants' examination of the stimulus set and their ruminations as to how to organize the stimuli. The critical question is how will SUSTAIN's two clusters be organized? Using the same parameters that were used in Billman and Knutson's (1996) studies (see Table 1 under the heading "Unsupervised"), SUSTAIN correctly predicted that the majority of sorts will be organized along one stimulus dimension. In particular, SUSTAIN predicted that 99% of sorts should be unidimensional and 1% of sorts should respect the intercorrelated structure of the stimulus set.

SUSTAIN's natural bias to focus on a subset of stimulus dimensions (which is further stressed by the selective attention mechanism) led it to predict the predominance of unidimensional sorts. Attention was directed toward stimulus dimensions that consistently matched at the cluster level. This led to certain dimensions becoming more salient over the course of learning (i.e., their λ attention value became larger). The dimension that developed the greatest saliency over the course of learning became the basis for the unidimensional sort. Thus, SUSTAIN predicts that the dimension a participant sorts the stimuli on is dependent on the order in which the participant encounters the stimuli. Of course, there are other possible explanations for why humans sort on a particular dimension (e.g., individual differences in a priori dimensional saliency). However, Gureckis and Love (2002) recently tested SUSTAIN's stimulus ordering prediction in a sequential sorting study, and human participants displayed the ordering result that SUSTAIN predicts.

It is interesting to note that SUSTAIN was able to account for both Billman and Knutson's (1996) data and Medin et al.'s (1987) data despite the differences in the findings. Participants in Billman and Knutson's studies infrequently organized the stimulus set along one dimension (especially in the intercorrelated conditions) because the correlations between dimensions were perfect. In contrast, each pairwise correlation in Medin et al. contained two exceptions (see Table 10). The perfect correlations in Billman and Knutson's studies led SUSTAIN to focus on a set of dimensions and not a single dimension.

The combined fits of Billman and Knutson's (1996) studies and Medin et al. (1987) suggest that the saliency of stimulus dimensions changes as a result of unsupervised learning and that the correlated structure of the world is most likely to be respected when there are numerous intercorrelated dimensions that are strong. Indeed, Younger and Cohen (1986) reported that even 10-month-old infants are sensitive to perfect correlations (see Gureckis & Love, in press, for SUSTAIN's account of the developmental trends).

SUSTAIN predicts that the intercorrelated structure of a stimulus set can also be discovered when the intercorrelations are imperfect (as in Medin et al., 1987) if the correlations are numerous. In cases where the total number of correlations is modest, and the correlations are weak and not interrelated, SUSTAIN predicts that stimuli will be organized along a single dimension.

Table 10
The Logical Structure of the Perceptual Dimensions in Medin et al. (1987)

Stimuli	
1 1 1 1	0 0 0 0
1 1 1 0	0 0 0 1
1 1 0 1	0 0 1 0
1 0 1 1	0 1 0 0
0 1 1 1	1 0 0 0

[5] This modification proved to be unnecessary because an unmodified version of SUSTAIN recruited two clusters in 99% of simulations.

General Discussion

SUSTAIN is motivated by a few simple principles yet can account for a wide range of data. SUSTAIN begins small and expands its architecture when the problem dictates it. SUSTAIN expands in response to surprising events (e.g., a prediction error in a supervised learning task or a stimulus that mismatches existing knowledge structures in an unsupervised learning task). SUSTAIN expands its architecture by adding a cluster that encodes the surprising event. Future events can then be understood in terms of the new cluster (as well as the existing clusters). When a surprising event does not occur, similar items are clustered together. Clusters that are activated by numerous stimuli serve as abstractions that can be continuously updated. This simple learning procedure allows SUSTAIN to infer a category's structure. The category substructure SUSTAIN uncovers is dictated not only by the structure of the world (i.e., the actual structure of the categories) but by the learning task or current goals. SUSTAIN acquires different knowledge structures depending on the current learning task (e.g., inference, classification, unsupervised learning, category construction, etc.). The data fits presented here suggest that SUSTAIN discovers category substructure in a manner close to how human learners do.

For example, SUSTAIN successfully fit the learning curves from Nosofsky, Gluck, et al.'s (1994) replication of Shepard et al.'s (1961) studies of classification learning by matching its complexity to that of the learning problem. SUSTAIN's solutions to the six problems were highly interpretable. Although these data suggest that item learning should always be more difficult than category learning, SUSTAIN was able to fit Medin et al.'s (1983) data on identification and category learning in which human learners displayed an identification learning advantage with distinctive stimuli. SUSTAIN modeled these data by capturing an interaction between learning task and stimulus type in which identification learning benefits more than category learning from distinctive stimuli. Again, SUSTAIN's solution was interpretable; distinctive stimuli reduced the benefit of abstraction and attenuated the effects of cluster competition (both of these factors favor identification learning relative to category learning). The simulations suggest an explanation for why experts are proficient at operating at more specific (i.e., lower) category levels—when stimuli are perceived as more distinct, the preferred level of categorization tends to shift toward a more specific level. SUSTAIN offers a mechanistic account (that is motivated by a few simple principles) for why this shift should occur.

Without altering its operation, SUSTAIN also was able to capture data (Yamauchi et al., 2002; Yamauchi & Markman, 1998) comparing human inference and classification learning. In particular, SUSTAIN correctly predicted that inference learning is better suited for linear category structures, whereas classification is best matched with nonlinear category structures. The knowledge structures SUSTAIN acquired (e.g., the prototype of each category in the case of inference learning with the linear category structure) allowed it to correctly predict that inference followed by classification is a more efficient task ordering than the reverse ordering. In the case of the linear category structure, there was a tangible benefit of abstraction (which contrasted with the detrimental effects of abstraction in the Medin et al., 1983, study).

Finally, SUSTAIN was able to account for human learning in a series of unsupervised learning tasks. SUSTAIN's clustering process allowed it to correctly predict that human learners favor intercorrelated category structures (Billman & Knutson, 1996). Without altering the parameter values, SUSTAIN was also able to account for studies in which humans sort intercorrelated stimuli along a single dimension (Medin et al., 1987). SUSTAIN resolves this apparent contradiction in terms of the nature (intercorrelated vs. nonintercorrelated), strength, and numerosity of the correlations.

Future Directions

One exciting avenue of future work is applying SUSTAIN to kindred areas of psychological research such as object recognition research. SUSTAIN may inform theories and models of object recognition. Tarr and Pinker (1989) argued that object recognition is viewpoint dependent. According to Tarr and Pinker, people represent an object as a collection of 2-D views, as opposed to representing an object as a structural description that includes the object's features and the spatial relations among the features (e.g., Biederman, 1987). Multiple-view theories and models bear a resemblance to exemplar categorization models in that abstraction occurs indirectly by storing many examples–views of a category–object. Poggio and Edelman's (1990) multiple-views model of object recognition interpolated among all views of an object observed when classifying a novel view. Like exemplar models, Poggio and Edelman's model is not very economical and stores every training example. SUSTAIN may offer a better approach. SUSTAIN only stores views when SUSTAIN makes an incorrect prediction. In this fashion, SUSTAIN may only need to recruit a few clusters to correctly identify an object from a novel view. Views that are very similar or vary on input dimensions that are not critical to identifying the object would share a common cluster. An object whose appearance varies greatly with a change in viewpoint (e.g., a car) would require more clusters (i.e., stored views) than an object whose appearance differs little across viewpoints (e.g., a basketball). Applying SUSTAIN to object recognition data could lead to novel predictions and would go far in integrating categorization and object recognition research.

Beyond its potential to improve our understanding of the related domains of categorization and object recognition, the ideas underlying SUSTAIN may successfully address a fundamental flaw in the exemplar and view-based frameworks. In these approaches, the notion of an exemplar (or view) is typically left undefined (cf. Logan, 1988). To illustrate the problem, consider a learner focusing on a chair while walking across a room. At every moment the learner is exposed to a slightly different image. The viewpoint is constantly changing and with it changes a number of the chair's properties (e.g., the visible features, albedo, etc.). After walking across the room, is one exemplar stored or are a million stored? Exemplar models do not address this fundamental question but SUSTAIN does. SUSTAIN only recruits a new cluster in response to a surprising event. In the above example, all the above information would be integrated into a single cluster unless something unexpected was encountered. SUSTAIN does not merely replace one problem (defining what an exemplar is) with another (defining what a cluster is) either. SUSTAIN specifies when and how clusters are formed and updated. SUSTAIN's clustering method may prove useful in understanding how humans individuate in general (cf. Barsalou, Huttenlocher, & Lamberts, 1998).

Future improvements in SUSTAIN will likely focus on its cluster recruitment strategy and its psychological underpinnings. The current recruitment strategy is somewhat idealized. To account for a broader range of learning data, mechanisms will probably need to be probabilistic and have the ability to remove and combine existing clusters. Fortunately, work in the machine learning literature (Aha, Kibler, & Albert, 1991; Bradshaw, 1987) suggests avenues for these future explorations.

Implications

Human category learning is influenced by the structure inherent in the world, but human learning is also flexible and adaptive. For example, human learning is affected by how the learner interacts with the stimuli, the learner's goals, and the nature of the stimuli. Humans can learn under either supervised or unsupervised conditions. Within these broad learning modes, important differences exist. For example, inference and classification learning are both supervised learning modes but they give rise to very different acquisition patterns. A key challenge for models of categorization is to show corresponding flexibility without losing the ability to predict patterns of performance.

The most distinctive contribution of SUSTAIN is that it addresses the different ways in which goals and tasks affect learning. In doing so, SUSTAIN extends the scope of categorization models. Although previous models have been able to account for supervised category learning at a fine level of detail, they have not demonstrated a corresponding breadth with respect to multiple learning procedures. SUSTAIN's fit of the data sets represents an existence proof that a greater range of findings can be addressed by models of categorization. We believe that this is an important step because a move toward a greater variety of conditions that affect learning is also a move toward greater realism and ecological validity.

In addition to extending the space of tasks and data that can be formally modeled, SUSTAIN provides an explanation of how these different tasks and data sets are interrelated. This is critical. As we explore new paradigms to gain additional perspectives on the nature of human categorization, there is a risk that the findings will fracture, leading to different theories and models for different tasks. SUSTAIN brings together a number of seemingly disparate tasks in a coherent manner using a single set of principles.

SUSTAIN's achievements are not at the cost of an overly complex formalism. SUSTAIN is a fairly idealized and simple model, motivated by a few interrelated principles. Its operation is straightforward—start simple and add complexity (i.e., clusters) as needed. The fact that these principles appear to be successful in accounting for otherwise counterintuitive data suggests that human categorization also favors starting simple and adding complexity as needed (see also Ahn & Medin, 1992).

References

Aha, D. W., Kibler, D., & Albert, M. K. (1991). Instance-based learning algorithms. *Machine Learning, 6,* 37.

Ahn, W. K., & Medin, D. L. (1992). A two-stage model of category construction. *Cognitive Science, 16,* 81–121.

Anderson, J. (1991). The adaptive nature of human categorization. *Psychological Review, 98,* 409–429.

Ash, T. (1989). Dynamic node creation in backpropagation networks. *Connection Science, 1,* 365–375.

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

Ashby, F., Queller, S., & Berretty, P. M. (1999). On the dominance of unidimensional rules in unsupervised categorization. *Perception & Psychophysics, 61,* 1178–1199.

Ashby, F. G., & Townsend, J. T. (1986). Varieties of perceptual independence. *Psychological Review, 93,* 154–179.

Azimi-Sadjadi, M. R., Sheedvash, S., & Trujillo, F. O. (1993). Recursive dynamic node creation in multilayer neural networks. *IEEE Transactions on Neural Networks, 4,* 242–256.

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

Barsalou, L. W. (1991). Deriving categories to achieve goals. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 27, pp. 1–64). San Diego, CA: Academic Press.

Barsalou, L. W., Huttenlocher, J., & Lamberts, K. (1998). Basing categorization on individuals and events. *Cognitive Psychology, 36,* 203–272.

Berlin, B., Breedlove, D. E., & Raven, P. (1972). General principles of classification and nomenclature in folk biology. *American Anthropologist, 75,* 214–242.

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review, 94,* 115–147.

Billman, D., & Knutson, J. (1996). Unsupervised concept learning and value systematicity: A complex whole aids learning the parts. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22,* 458–475.

Bradshaw, G. (1987). Learning about speech sounds: The NEXUS project. In P. Langley (Ed.), *Proceedings of the Fourth International Workshop on Machine Learning* (pp. 1–11). Irvine, CA: Morgan Kaufmann.

Bruner, J. S., Goodnow, J. J., & Austin, G. A. (1956). *A study of thinking.* New York: Wiley.

Busemeyer, J., & McDaniel, M. A. (1997). The abstraction of intervening concepts from experience with multiple input-multiple output causal environments. *Cognitive Psychology, 32,* 1–48.

Carpenter, G. A., & Grossberg, S. (1987). A massively parallel architecture for a self-organizing neural pattern recognition machine. *Computer Vision, Graphics, and Image Proceedings, 37,* 54–115.

Carpenter, G. A., Grossberg, S., & Reynolds, J. H. (1991). ARTMAP: Supervised real-time learning and classification of nonstationary data by a self-organizing neural network. *Neural Networks, 4,* 565–588.

Chin-Parker, S., & Ross, B. H. (2002). The effect of category learning on sensitivity to within category correlations. *Memory & Cognition, 30,* 353–362.

Cho, S. (1997). Self-organizing map with dynamical node splitting: Application to handwritten digit recognition. *Neural Computation, 9,* 1345–1355.

Clapper, J. P., & Bower, G. H. (1991). Learning and applying category knowledge in unsupervised domains. *Psychology of Learning and Motivation, 27,* 65–108.

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

Diamond, R., & Carey, S. (1986). Why faces are and are not special: An effect of expertise. *Journal of Experimental Psychology: General, 115,* 107–117.

Elman, J. L. (1994). Implicit learning in neural networks: The importance of starting small. In C. Umilta & M. Moscovitch (Eds.), *Attention and performance XV: Conscious and nonconscious information processing* (pp. 861–888). Cambridge, MA: MIT Press.

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

Fahlman, S. E., & Lebiere, C. (1990). The cascade-correlation learning architecture. In D. S. Touretzky (Ed.), *Advances in neural information processing systems 2. Proceedings of the 1989 conference* (pp. 524–532). San Mateo, CA: Morgan Kaufmann.

Farah, M. J. (1992). Is an object an object an object? Cognitive and neuropsychological investigations of domain specificity in visual object recognition. *Current Directions in Psychological Science, 1,* 164–169.

Feldman, J. (2000, October 5). Minimization of boolean complexity in human concept learning. *Nature, 407,* 630–633.

Garner, W. R., & Whitman, J. (1965). Form and amount of internal structure as factors in free-recall learning of nonsense words. *Journal of Verbal Learning and Verbal Behavior, 4,* 257–266.

Gauthier, I., & Tarr, M. J. (1997). Becoming a "Greeble" expert: Exploring the face recognition mechanism. *Vision Research, 37,* 1673–1682.

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural Computation, 4,* 1–58.

Gluck, M. A., & Bower, G. H. (1988). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General, 117,* 225–244.

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

Goldstone, R. L., Schyns, P. G., & Medin, D. L. (1997). Learning to bridge between perception and cognition. In R. L. Goldstone, P. G. Schyns, & D. L. Medin (Eds.), *The psychology of learning and motivation* (Vol. 36, pp. 1–14). San Diego, CA: Academic Press.

Gosselin, F., & Schyns, P. (2001). Why do we SLIP to the basic-level? Computational constraints and their implementation. *Psychological Review, 108,* 735–758.

Gureckis, T., & Love, B. C. (2002). Who says models can only do what you tell them? Unsupervised category learning data, fits, and predictions. In W. D. Gray & C. D. Schunn (Eds.), *Proceedings of the 24th Annual Conference of the Cognitive Science Society* (pp. 399–404). Hillsdale, NJ: Erlbaum.

Gureckis, T., & Love, B. C. (2003). Towards a unified account of supervised and unsupervised learning. *Journal of Experimental and Theoretical Artificial Intelligence, 15,* 1–14.

Gureckis, T., & Love, B. C. (in press). Common mechanisms in infant and adult category learning. *Infancy.*

Hartigan, J. A. (1975). *Clustering algorithms.* New York: Wiley.

Heit, E. (1992). Categorization using chains of examples. *Cognitive Psychology, 24,* 341–380.

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

Hintzman, D. L. (1986). Schema abstraction in a multiple-trace memory model. *Psychological Review, 93,* 411–428.

Homa, D., & Cultice, J. (1984). Role of feedback, category size, and stimulus distortion on the acquisition and utilization of ill-defined categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10,* 83–94.

Hovland, C. L., & Weiss, W. (1953). Transmission of information concerning concepts through positive and negative information. *Journal of Experimental Psychology, 45,* 175–182.

Johnson, K. E., & Mervis, C. B. (1997). Effects of varying levels of expertise on the basic level of categorization. *Journal of Experimental Psychology: General, 126,* 248–277.

Karnin, E. D. (1990). A simple procedure for pruning back-propagation trained neural networks. *IEEE Transactions on Neural Networks, 1,* 239–242.

Kohonen, T. (1982). Self-organizing formation of topologically correct feature maps. *Biological Cybernetics, 43,* 59–69.

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

Kruschke, J. K. (1993). Human category learning: Implications for back propagation models. *Connection Science, 5,* 3–36.

Kruschke, J. K., & Movellan, J. R. (1991). Benefits of gain: Speeding learning and minimal hidden layers in back-propagation networks. *IEEE Transactions on Systems, Man, and Cybernetics, 99,* 21.

Lassaline, M. E. (1996). Structural alignment in induction and similarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22,* 754–770.

Lee, M. D., & Navarro, D. J. (2002). Extending the ALCOVE model of category learning to featural stimulus domains. *Psychonomic Bulletin & Review, 9,* 43–58.

Levine, D. (1996). *Users guide to the pgapack parallel genetic algorithm library* (Tech. Rep. No. ANL-95/18). Argonne, IL: Argonne National Laboratory.

Logan, G. D. (1988). Towards an instance theory of automatization. *Psychological Review, 95,* 492–527.

Lopez, A., Atran, S., Coley, J. D., Medin, D. L., & Smith, E. E. (1997). The tree of life: Universal and cultural features of folk biological taxonomies and inductions. *Cognitive Psychology, 32,* 251–295.

Love, B. C. (2000). Learning at different levels of abstraction. In L. R. Gleitman & A. K. Joshi (Eds.), *Proceedings of the 22nd Annual Conference of the Cognitive Science Society* (pp. 800–805). Mahwah, NJ: Erlbaum.

Love, B. C. (2001). Three deadly sins of category learning modelers. *Behavioral and Brain Sciences, 24,* 687–688.

Love, B. C. (2002). Comparing supervised and unsupervised category learning. *Psychonomic Bulletin & Review, 9,* 829–835.

Love, B. C. (2003). The multifaceted nature of unsupervised category learning. *Psychonomic Bulletin & Review, 10,* 190–197.

Love, B. C., Markman, A. B., & Yamauchi, T. (2000). Modeling classification and inference learning. In H. Kautz & B. Porter (Eds.), *Proceedings of the Seventeenth National Conference on Artificial Intelligence* (pp. 136–141). Cambridge, MA: MIT Press.

Love, B. C., & Medin, D. L. (1998a). Modeling item and category learning. In M. A. Gernsbacher & S. J. Derry (Eds.), *Proceedings of the 20th Annual Conference of the Cognitive Science Society* (pp. 639–644). Mahwah, NJ: Erlbaum.

Love, B. C., & Medin, D. L. (1998b). SUSTAIN: A model of human category learning. In C. Rich & J. Mostow (Eds.), *Proceedings of the Fifteenth National Conference on Artificial Intelligence* (pp. 671–676). Cambridge, MA: MIT Press.

Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis.* Westport, CT: Greenwood Press.

Maddox, W. T., & Ashby, F. G. (1993). Comparing decision bound and exemplar models of categorization. *Perception & Psychophysics, 53,* 49–70.

Malt, B. C. (1995). Category coherence in cross-cultural perspective. *Cognitive Psychology, 29,* 85–148.

Malt, B. C., Murphy, G. L., & Ross, B. H. (1995). Predicting features for members of natural categories when categorization is uncertain. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 646–661.

Markman, A. B., & Makin, V. S. (1998). Referential communication and category acquisition. *Journal of Experimental Psychology: General, 127,* 331–354.

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

Marr, D. (1982). *Vision.* San Francisco: Freeman.

McKinley, S. C., & Nosofsky, R. M. (1996). Selective attention and the formation of linear decision boundaries. *Journal of Experimental Psychology: Human Perception and Performance, 22,* 294–317.

Medin, D. L., & Bettger, J. G. (1994). Presentation order and recognition

of categorically related examples. *Psychonomic Bulletin & Review, 1,* 250–254.

Medin, D. L., Dewey, G. I., & Murphy, T. D. (1983). Relationships between item and category learning: Evidence that abstraction is not automatic. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 9,* 607–625.

Medin, D. L., Lynch, E. B., Coley, J. D., & Atran, S. (1997). Categorization and reasoning among tree experts: Do all roads lead to Rome? *Cognitive Psychology, 32,* 49–96.

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

Medin, D. L., & Shoben, E. J. (1988). Context and structure in conceptual combination. *Cognitive Psychology, 20,* 158–190.

Medin, D. L., Wattenmaker, W. D., & Hampson, S. E. (1987). Family resemblance, conceptual cohesiveness, and category construction. *Cognitive Psychology, 19,* 242–279.

Minsky, M. L., & Papert, S. (1969). *Perceptrons: An introduction to computational geometry.* Cambridge, MA: MIT Press.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review, 92,* 289–316.

Murphy, G. L., & Ross, B. H. (1994). Predictions from uncertain categorizations. *Cognitive Psychology, 27,* 148–193.

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

Nosofsky, R. M. (1991a). Relation between the rational model and the context model of categorization. *Psychological Science, 2,* 416–421.

Nosofsky, R. M. (1991b). Tests of an exemplar model for relating perceptual classification and recognition memory. *Journal of Experimental Psychology: Human Perception and Performance, 17,* 3–27.

Nosofsky, R. M., Gluck, M. A., Palmeri, T. J., McKinley, S. C., & Glauthier, P. (1994). Comparing models of rule based classification learning: A replication and extension of Shepard, Hovland, and Jenkins (1961). *Memory & Cognition, 22,* 352–369.

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., Lopez, A., & Shafir, E. (1990). Category-based induction. *Psychological Review, 97,* 185–200.

Poggio, T., & Edelman, S. (1990, January 18). A network that learns to recognize three-dimensional object. *Nature, 343,* 263–266.

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology, 77,* 241–248.

Rhodes, G., Tan, S., Brake, S., & Taylor, K. (1989). Expertise and configural coding in face recognition. *British Journal of Psychology, 80,* 313–331.

Rips, L. J. (1975). Inductive judgments about natural categories. *Journal of Verbal Learning and Verbal Behavior, 14,* 665–681.

Rosch, E. (1975). Universals and cultural specifics in human categorization. In R. Brislin, S. Bochner, & W. Lanner (Eds.), *Cross-cultural perspectives on learning* (pp. 177–206). New York: Wiley.

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology, 7,* 573–605.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology, 8,* 382–439.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage in the brain. *Psychological Review, 65,* 386–408.

Ross, B. H. (1996). Category representations and the effects of interacting with instances. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22,* 1249–1265.

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

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature, 323,* 533–536.

Schank, R. C., Collins, G. C., & Hunter, L. E. (1986). Transcending inductive category formation in learning. *Behavioral and Brain Sciences, 9,* 639–686.

Shepard, R. N. (1987, September 11). Toward a universal law of generalization for psychological science. *Science, 237,* 1317–1323.

Shepard, R. N., Hovland, C. L., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs, 75*(13, Whole No. 517).

Sloman, S. A. (1997). Explanatory coherence and the induction of properties. *Thinking and Reasoning, 3,* 81–110.

Solomon, K. O., Medin, D. L., & Lynch, E. B. (1999). Concepts do more than categorize. *Trends in Cognitive Science, 3,* 99–105.

Tanaka, J. W., & Taylor, M. (1991). Object categories and expertise: Is the basic level in the eye of the beholder? *Cognitive Psychology, 23,* 457–482.

Tarr, M. J., & Pinker, S. (1989). Mental rotation and orientation-dependence in shape recognition. *Cognitive Psychology, 21,* 233–282.

Waxman, S. R. (1998). Linking object categorization and naming: Early expectations and the shaping role of language. In D. L. Medin (Ed.), *The psychology of learning and motivation* (Vol. 38, pp. 249–291). San Diego, CA: Academic Press.

Widrow, B., & Hoff, M. E. (1960). Adaptive switching circuits. In *Western Electronic Show and Convention: Convention record* (Vol. 4, pp. 96–104). New York: Institute of Radio Engineers.

Yamauchi, T., Love, B. C., & Markman, A. B. (2002). Learning nonlinearly separable categories by inference and classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28,* 585–593.

Yamauchi, T., & Markman, A. B. (1998). Category learning by inference and classification. *Journal of Memory and Language, 39,* 124–149.

Yamauchi, T., & Markman, A. B. (2000). Inference using categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 776–795.

Younger, B., & Cohen, L. B. (1986). Developmental change in infants' perception of correlations among attributes. *Child Development, 57,* 803–815.

(*Appendix follows*)

Appendix

Details on ALCOVE

Although ALCOVE and SUSTAIN differ in a number of critical ways, their formal descriptions overlap extensively. The description of ALCOVE in this appendix is leveraged off of SUSTAIN's description. We recommend reading the sections SUSTAIN's Formalization and *Exemplar-Based Approaches* before reading this appendix.

Readers familiar with ALCOVE might notice some minor differences with other implementations. To make ALCOVE more psychologically realistic and comparable with SUSTAIN, exemplar nodes are added as they are encountered instead of the entire set being instantiated on the first trial of a simulation. Also, weights are updated on a trial-by-trial basis instead of by block (i.e., batch updating).

Stimulus Representation

The original ALCOVE can represent continuous and binary-valued dimensions, but not nominally valued dimensions. Later versions introduce the ability to represent features (Lee & Navarro, 2002). To allow ALCOVE to represent nominally valued dimensions and to facilitate comparisons with SUSTAIN, stimuli are represented in a manner identical to how they are represented in SUSTAIN. Distance in representational space between a stimulus and an exemplar is calculated the same way as distance between a stimulus and a cluster is calculated in SUSTAIN. In ALCOVE, an exemplar is recruited for each novel stimulus.

ALCOVE's Parameters

Table A1 lists all of ALCOVE's parameters and includes a brief description of the function of each parameter. The best fitting values for each study are also shown. Differences exist in how SUSTAIN and ALCOVE are parameterized. ALCOVE contains two learning rates, one for learning how to shift attention and another for learning association weights between exemplars and category units (i.e., responses), whereas SUSTAIN contains just one learning rate. SUSTAIN's learning rate does have the attentional focus parameter *r*, which affects how quickly attention shifts, so SUSTAIN also has a mechanism for selectively adjusting the speed of attentional shifts.

Unlike SUSTAIN, ALCOVE does not contain a cluster- (or exemplar) competition parameter. ALCOVE does have a specificity parameter that is somewhat related to SUSTAIN's cluster-competition parameter. The specificity parameter governs how strongly exemplars will be activated that are not identical to the current stimulus. To the extent that other exemplars can be understood as competing with the dominant exemplar, this parameter is analogous to SUSTAIN's cluster-competition parameter. Of course, exemplars can also behave cooperatively as in cases in which exemplar models display prototype effects (strong endorsement of the underlying prototype) because of the cumulative similarity of the prototype stimulus to all stored exemplars (see Medin & Schaffer, 1978). For data sets in which the saliency of a stimulus dimension is not controlled, the stimulus representation is parameterized in the same fashion as it was with SUSTAIN.

Mathematical Formulation of ALCOVE

In ALCOVE, all hidden units (i.e., exemplars) have a nonzero output. In SUSTAIN, only one hidden unit (i.e., cluster) has a nonzero output. The output of an exemplar is:

$$H_j^{\text{out}} = e^{-c(\sum_{i=1}^{m} \lambda_i \mu_{ij})}, \tag{A1}$$

where H_j^{out} is the activation of the jth exemplar, c is the specificity parameter, m is the number of stimulus dimensions, λ_i is the attention weight for the ith input dimension, and μ_{ij} is the distance between exemplar H_j's position in the ith dimension and the current stimulus's position in the ith dimension (μ_{ij} is defined as it is in SUSTAIN).

Activation is spread from exemplars to the output units forming the queried stimulus dimension in the exact same fashion as SUSTAIN passes activation from clusters to output units (see Equation 7). The decision probabilities are also calculated in the same fashion (see Equation 8).

After responding, target values for learning are calculated as they are in SUSTAIN (see Equation 9). Unlike SUSTAIN, hidden units (i.e., exemplars) in ALCOVE do not shift their positions as clusters do in SUSTAIN. Like SUSTAIN, ALCOVE reallocates attention after receiving feedback:

$$\Delta\lambda_i = -\eta_a \sum_{j=1}^{n} \Big[\sum_{k=1}^{v_z} (t_{zk} - C_{zk}^{\text{out}}) w_{j,zk} H_j^{\text{out}} \cdot c \cdot \mu_{ij}, \tag{A2}$$

where $\Delta\lambda_i$ is the change in attention for Dimension i, n_a is the attention learning rate, n is the number of hidden units, z is the queried dimension, and v_z is the number of nominal values for Dimension z. ALCOVE's attentional mechanism seeks to minimize overall error and is derived from the delta learning rule (Rumelhart et al., 1986). Initially, the attention weight for each dimension is set to 1.

As in SUSTAIN, the one-layer delta learning rule (Widrow & Hoff, 1960) is used to adjust the weights between hidden units (exemplars) and the category units forming the queried dimension, *z:*

$$\Delta w_{j,zk} = \eta_w (t_{zk} - C_{zk}^{\text{out}}) H_j^{\text{out}}. \tag{A3}$$

Table A1
ALCOVE's Best Fitting Parameters for All Data Sets Considered

Function/adjusts	Symbol	Six types	First/last name	Infer./class.	Unsupervised
Specificity	c	6.067453	2.106726	8.359079	6.716753
Decision consistency	d	3.803207	13.580068	3.782043	3.50684843
Attention learning rate	η_a	0.02039906	0.361167	0.3869768	0.005
Weight learning rate	η_w	0.1132104	0.058514	0.09300784	3.50684843
Category focus	λ_{label}	—	—	10.46671	—
Distinct focus	$\lambda_{\text{distinct}}$	—	1.652672	—	—

Note. Dashes indicate that the parameter was not applicable to the simulation. Infer. = inference; class. = classification.

Received May 30, 2001
Revision received June 27, 2003
Accepted June 27, 2003 ■