Towards a unified account of supervised and unsupervised category learning

TODD M. GURECKIS and BRADLEY C. LOVE

Department of Psychology, SEA 4.310J, University of Texas at Austin, Austin, TX 78712, USA
e-mail: gureckis@love.psy.utexas.edu; love@love.psy.utexas.edu

Abstract. (Supervised and Unsupervised STratified Adaptive IncrementalNetwork) is a network model of human category learning. 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 has expanded the scope of findings that models of human category learning can address. This paper extends SUSTAIN so that it can be used to account for both supervised and unsupervised learning data through a common mechanism. A modified recruitment rule is introduced that creates new conceptual clusters in response to surprising events during learning. The new formulation of the model is called uSUSTAIN for 'unified SUSTAIN.' The implications of using a unified recruitment method for both supervised and unsupervised learning are discussed.

1. Introduction

Categorization plays a central role in the cognitive ability of humans. Our ability to reason, make decisions, recognize objects and process language all depend on being able to organize knowledge about the environment into categories. Work in understanding human categorization behaviour has traditionally focused on studying human performance in supervised learning tasks. The typical experimental procedure used for studying this type of learning involves asking a participant to learn to classify a set of stimuli while receiving corrective feedback on every trial (Shepard *et al.* 1961, Nosofsky *et al.* 1994a). The use of this common approach has facilitated comparisons across studies.

However, this kind of 'teacher-guided' learning may account for only a small portion of the learning that we are engaged in each day. A large part of our learning is better characterized as *unsupervised* because no explicit feedback is available from the environment. For example, we often categorize incoming email as belonging to

the 'junk mail' category or to the 'interesting mail' category. We are not explicitly taught to identify members of either category and we do not receive specific feedback on each example. Nevertheless, we acquire and use categories such as these to sort our mail on a daily basis.

It is clear that humans can acquire and utilize information in a variety of ways, but how different are the cognitive processes underlying these ways of learning? In this paper, we attempt to better define the relationship between supervised and unsupervised learning by demonstrating how a single modelling framework can account for a diverse set of findings from both the supervised and unsupervised category learning literatures. Our results suggest that these two types of learning may, in fact, be quite similar.

The question of how people learn from their environment in the absence of explicit feedback has largely been ignored by psychologists studying category learning (see Clapper and Bower 1994, Billman and Knutson 1996, Ashby *et al.* 1999, for some exceptions). A contributing reason for this is the assumption that supervised and unsupervised learning are quite different processes with separate goals and under-lying mechanisms. For example, supervised learning is usually characterized as intentional, in that learners actively search for rules (perhaps by hypothesis testing) and are explicitly aware of the rule they are considering (Nosofsky *et al.* 1994b). On the other hand, unsupervised learning is seen as incidental, in that the criteria for category membership is not usually available to the learner in an explicit sense. In line with this position is the view that unsupervised learning is an undirected, stimulus driven, incremental accrual of information (Hock *et al.* 1986, Lewicki 1986, Hayes and Broadbent 1988, Berry and Dienes 1993, Cleermans 1993).

Recently, however, these assumptions about unsupervised learning have proven incorrect. Love (2002b) has shown that unsupervised learning can take a variety of forms in which successful learning depends on an advantageous pairing of the structure of the learning problem and the manner in which the subject interacts with the stimuli (i.e. unsupervised induction task). This result parallels findings in supervised learning (Yamauchi *et al.* 2002). In addition, Love (2002a) found that performance in unsupervised learning under intentional conditions (i.e. when participants are aware of the learning task) is very similar to performance in supervised classification learning. These two results suggest that supervised and unsupervised learning may be less conceptually distinct than once assumed.

Few previous modelling attempts have made much of an effort to account for human learning performance in tasks other than supervised learning (see Billman and Heit 1988, Clapper and Bower 1991, for exceptions) and even less work has been done in modelling both supervised and unsupervised learning within a common framework. A notable exception is Anderson's (1991) rational model. The rational model operates within a Bayesian framework and attempts to uncover the covert category structures that are maximally informative in terms of predictive inference (e.g. knowing about one aspect of stimulus allows for the other aspects to be inferred). The rational model formally unites supervised and unsupervised category learning (the category label is treated the same as other aspects of the stimulus), but it does not make allowances for the fact that human category learning is largely driven by current goals and the nature of the induction task. The rational model's abstract, information-based approach to category learning makes it impossible for it to address key human category learning studies (Love *et al.* 2002).

1.1. *Organization of Paper*

In this paper, we present a new modelling approach that unifies both supervised and unsupervised learning under a single principle of adaptation to surprise. We begin with a psychological model of category learning called SUSTAIN (Supervised and Unsupervised STratified Adaptive Incremental Network) and alter its operation to address both forms of learning. The new version of the model is referred to as uSUSTAIN for 'unified SUSTAIN.' Unlike the rational model, uSUSTAIN is sensitive to learning goals and to the nature of the induction task. We then evaluate the success of this modelling approach by examining uSUSTAIN's fit of a diverse set of human category learning studies.

This paper is organized as follows: first, we describe our modelling approach based on SUSTAIN. Next, we overview the human category learning data sets that we have considered and examine uSUSTAIN's ability to account for this data. We conclude by summarizing the findings and implications of our unified account of supervised and unsupervised learning.

2. Modeling approach: an introduction to SUSTAIN and uSUSTAIN

SUSTAIN is a network model of human category learning. The model has been successfully applied to an array of challenging human data sets spanning a variety of category learning paradigms including classification learning (Love and Medin 1998b), learning at different levels of abstraction (Love *et al.* 2000), and inference learning (Love and Medin 1998a). We begin our introduction to SUSTAIN by presenting a overview of the operation of the model. This is followed by a discussion of the key psychological principles from which the model is derived. This introduction serves to highlight the important features of the model and provides the motivation for the later sections. This general introduction to the model is followed by a section which explains the mathematical equations that follow from SUSTAIN's general principles. We conclude the description of our modelling approach with a discussion of the challenges of modelling both supervised and unsupervised learning within a single framework. In this section, we describe how SUSTAIN may be modified to use a flexible and intuitive notion of *surprise* to provide an unified account of these two types of learning.

2.1. *Overview of model*

SUSTAIN is a clustering model of human category learning. The model takes as input a set of perceptual features that are organized into a series of independent feature dimensions. The internal representations in the model consist of a set of clusters. Categories are represented in the model as one or more associated clusters. Initially, the network only has only one cluster that is centred upon the first input pattern. As new stimulus items are presented, the model attempts to assign new items to an existing cluster. This assignment is done through an unsupervised procedure based on the similarity of the new item to the stored clusters. When a new item is assigned to a cluster, this cluster updates its internal representation to become the average of all items assigned to the cluster so far. However, if SUSTAIN discovers through feedback that this similarity based assignment is incorrect, a new cluster is created to encode the exception. Classification decisions are ultimately based on the cluster to which an instance is assigned.

2.2. *The key principles of SUSTAIN*

With this general understanding of the operation of the model in mind, we now examine the five key principles of SUSTAIN. These principles highlight the important features of the model and provide the foundation for the model's formalism.

2.2.1. *Principle 1, SUSTAIN is biased towards simple solutions*. SUSTAIN is initially directed towards simple solutions. At the start of learning, SUSTAIN has only one cluster which is centreed on the first input item. It then adds clusters (i.e. complexity) only as needed to accurately describe the category structure. Like other models of category learning (e.g. Kruschke, 1992), SUSTAIN maintains an attentional tuning mechanism which allows it to selectively weight stimulus feature dimensions. During the process of learning, SUSTAIN updates these attentional weights to place emphasis on stimulus dimensions that are most useful for categorization. Its selective attention mechanism further serves to bias SUSTAIN towards simple solutions by focusing the model on the stimulus dimensions that provide consistent information.

2.2.2. *Principle 2, similar stimulus items tend to cluster together*. In learning to classify stimuli as members of two distinct categories, SUSTAIN will cluster similar items together. For example, different instances of a bird subtype (e.g. sparrows) could cluster together and form a sparrow cluster instead of leaving separate traces in memory for each instance. Clustering is an unsupervised process because cluster assignment is done on the basis of similarity, not feedback.

2.2.3. *Principle 3, SUSTAIN learns in both a supervised and unsupervised fashion*. In learning to classify the categories 'birds' and 'mammals', SUSTAIN relies on both unsupervised and supervised learning processes. Consider a learning trial in which SUSTAIN has formed a cluster whose members are small birds, and another cluster whose members are four-legged mammals. If SUSTAIN is subsequently asked to classify a bat, it will initially predict that a bat is a bird on the basis of overall similarity (bats and birds are both small, have wings, fly, etc.). Upon receiving feedback from the environment (supervision) indicating that a bat is a mammal, SUSTAIN will recruit a new cluster to represent the bat as an exception to the mammal category. The next time SUSTAIN is exposed to the bat or another similar bat, SUSTAIN will correctly predict that a bat is a mammal. This example also illustrates how SUSTAIN can entertain more complex solutions when necessary through cluster recruitment (see Principle 1).

2.2.4. *Principle 4, the pattern of feedback matters*. As the example used above illustrates, feedback affects the inferred category structure. Prediction failures result in a cluster being recruited, thus different patterns of feedback can lead to different representations being acquired. This principle allows SUSTAIN to predict different acquisition patterns for different learning modes (e.g. inference versus classification learning) that are informationally equivalent but differ in their pattern of feedback.

2.2.5. *Principle 5, cluster competition*. Clusters can be seen as competing explanations of the input. The strength of the response from 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 (see Sloman's [1997], account of competing explanations in reasoning).

2.3. *Mathematical formulation of SUSTAIN*
This section of the paper explains how the general principles that govern SUS-TAIN's operation are implemented in an algorithmic model.

2.3.1. *Input representation.* Stimuli are represented in the model 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 colour) and are members of one of two categories would require a vector frame with four dimensions. A four dimensional binary-valued stimulus (three perceptual dimensions plus the category label) can be thought of as a four character string (e.g. **1 2 1 1**) in which each character represents the value of a stimulus dimension. For example, the first character could denote the size dimension with a **1** indicating a small stimulus and a **2** indicating a large stimulus.

Of course, a learning trial usually involves an incomplete stimulus representation. For instance, 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, subjects 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 **? 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).

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.

In order 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 utilized. 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 $I^{pos_{ik}}$ 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 $I^{pos_{32}}$ would equal one, whereas $I^{pos_{31}}$ and $I^{pos_{33}}$ would equal zero. The 'pos' in I^{pos} denotes that the current stimulus is located at a particular position in a multi-dimensional representational space.
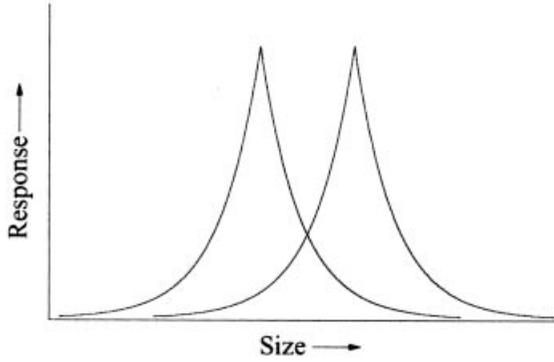
Figure 1. Two receptive fields are shown. These receptive fields are for the same
dimension (i.e. size) and accordingly have the same tuning, but are centred at
different positions along the dimension. The cluster containing the receptive
field on the right prefers larger stimuli than the cluster containing the
receptive field on the left.

2.3.2. *Receptive fields*. Each cluster has a receptive field for each stimulus
dimension. A cluster's receptive field for a given dimension is centreed at the
cluster's position along that dimension. The position of a cluster within a
dimension indicates the cluster's expectations for its members. Figure 1 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. 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 centre 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
centre of the field, and $\alpha(\mu)$ denotes the response of the receptive field to a stimulus
falling μ units from the centre 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 λ 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:
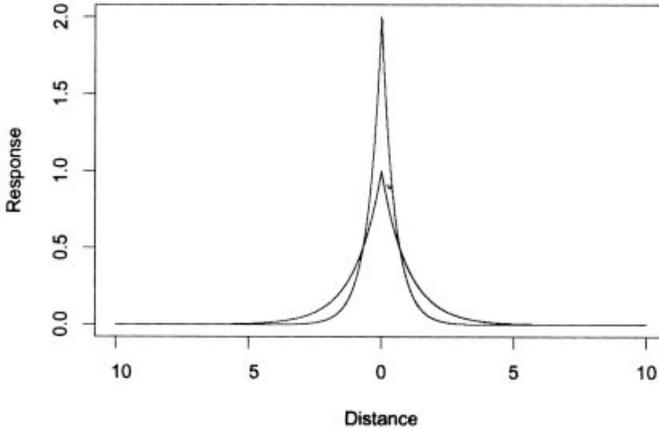
Figure 2. Two receptive fields are shown. A maximal response is elicited from both receptive fields when a stimulus falls in the centre of each receptive field (a 1.0 response for the broadly tuned receptive field; a 2.0 response for the tightly tuned field). Compared to the broadly tuned field, the tightly tuned field's response is stronger for stimuli falling close to the centre and is weaker for stimuli farther from the centre. The crossover point occurs at a distance from centre of 0.7 (approximately).

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

These properties of exponentials prove useful in formulating SUSTAIN.

2.3.3. *Cluster activation.* With nominal stimulus dimensions, the distance μ_{ij} (from 0 to 1) between the ith dimension of the stimulus and cluster j's position along the ith dimension is:

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

where v_i is the number of different nominal values on the ith dimension, I is the input representation (as described in a previous section), and $H_j^{pos_{ik}}$ is cluster j's position on the ith dimension for value k (the sum of all k 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. For example, a cluster in which 20% of the members are single, 45% are married, and 35% are divorced will converge to the location [0.20 0.45 0.35] 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:

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

where H_j^{act} is the activation of the jth cluster, m is the number of stimulus dimensions, λ_i is the tuning of the receptive field for the ith input dimension, and

r is an attentional parameter (always non-negative). 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 will have 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 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.

2.3.4. *Competition*. 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}

$$H_j^{out} = \frac{(H_j^{act})^{\beta}}{\sum_{i=1}^{n}(H_i^{act})^{\beta}} H_j^{act} \tag{6}$$

For all other H_j

$$H_j^{out} = 0$$

where n is the number of clusters and β is the lateral inhibition parameter (always non-negative) that regulates cluster competition. When β is small, competing clusters strongly inhibit the winner. When β is large the winner is weakly inhibited. Clusters other than the winner 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 will reduce confidence in a choice (reflected in a lower output value).

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

$$C_{zk}^{out} = \sum_{j=1}^{n} w_{j,zk} H_j^{out} \tag{7}$$

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 did 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:

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

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 choice rule is conceptually related to this decision rule (Luce, 1959).

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

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

Kruschke (1992) refers to this kind of teaching signal as a 'humble teacher' and explains 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*},
> then recruit a new cluster. $\tag{10}$

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.

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

2.3.6. *Learning.* The position of the winner is adjusted:

> For the winning H_j,
> $$\Delta H_j^{pos_{ik}} = \eta(I^{pos_{ik}} - H_j^{pos_{ik}}) \tag{11}$$

where η is the learning rate. The centres of the winner's receptive fields move towards the input pattern according to the Kohonen learning rule. This learning rule centres the cluster amidst its members.

Using our result from equation (3), receptive field tunings are updated according to:

$$\Delta \lambda_i = \eta e^{-\lambda_i \mu_{ij}} \left(1 - \lambda_i \mu_{ij} \right) \tag{12}$$

where j is the index of the winning cluster.

Only the winning cluster updates the value of λ_i. Equation (12) 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 of λ_i for this case is 1 (i.e. equation (12) 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 and Hoff, 1960) is used to adjust these weights:

$$\Delta w_{j,zk} = \eta(t_{zk} - C_{zk}^{out})H_j^{out} \tag{13}$$

where z is the queried dimension. Note that only the winning cluster will have its weights adjusted since it is the only cluster with a non-zero output.

2.4.   *uSUSTAIN: a unified approach to supervised and unsupervised learning*

In the formulation of SUSTAIN described above, the network adapts its architecture in response to external feedback. Only when SUSTAIN predicts an incorrect response does it recruit a new cluster to capture the exception. Thus, SUSTAIN changes its architecture in response to a *surprising* event, which in this case is a misclassified item. Unfortunately, this recruitment rule leaves SUSTAIN unable to model unsupervised learning data. In unsupervised learning, there is no feedback and we assume that each stimulus item is a member of the same category (the global category). SUSTAIN's supervised recruitment process is disabled because prediction errors do not occur in unsupervised learning.

In previous work (Love *et al.* 2002), we augmented SUSTAIN with a second recruitment rule for unsupervised learning situations. In this work, SUSTAIN recruited a new cluster when the current stimulus item was not sufficiently similar to any existing cluster. Like its supervised analog, this unsupervised recruitment rule is based on a notion of surprise. In the case of unsupervised recruitment, SUSTAIN recruits a new cluster in response to a surprisingly novel or unfamiliar stimulus item. Under both of these recruitment procedures, a new cluster is added when the existing clusters do not properly characterize the stimulus. Although the two separate recruitment procedures have been successful, one unified recruitment procedure would be preferable. Beyond parsimony, a unified account could prove useful in clarifying the relationship between unsupervised and supervised learning.

A simple way to integrate the two recruitment strategies is to generalize the unsupervised procedure so that it is applicable to supervised learning situations. Under this scheme, a new cluster is recruited when the current stimulus is not sufficiently similar to any cluster in its category:

$$\text{if } (A_{H_j} < \tau), \text{ then recruit a new cluster} \tag{14}$$

where A_{H_j} is the activation of the most highly activated cluster that belongs to the same category as the current input stimulus and τ is a constant between 0 and 1 (a parameter). In unsupervised learning, all items belong to the same category, thus A_{H_j} refers to the most activated cluster overall. In supervised learning, the most activated cluster predicting the correct category may not in fact be the most activated cluster overall.

Besides providing a unified framework, this recruitment strategy has a number of other virtues over SUSTAIN's original recruitment rule for supervised learning. For example, the unified procedure will recruit a new cluster when an unusual item is encountered that does not result in a prediction error whereas the previous error-

driven recruitment scheme would not recruit a new cluster to encode the unusual item. Assigning a very unusual item to an existing cluster (a cluster the item is not very similar to) could result in catastrophic interference (see Ratcliff 1990) as the cluster must undergo radical change to accommodate its newest member. Still, it remains to be seen whether such a unified rule can provide an adequate account of both unsupervised and supervised learning. The following sections evaluate this possibility.

3. Modelling unsupervised learning

To evaluate the promise of uSUSTAIN in the domain of unsupervised learning, we provide results of it's application to Experiments 2 and 3 from Billman and Knutson's (1996) unsupervised learning study and to unsupervised category construction (i.e. sorting) data from Medin *et al.* (1987).

3.1. *Modelling Billman and Knutson (1996)*

Billman and Knutson's experiments tested the prediction that category learning is easier when certain stimulus attributes are predictive of other attributes by way of a correlation (e.g. 'has wings', 'can fly', 'has feathers' are all correlated features of birds). Their studies evaluate how relations among stimulus attributes affect learning in an unsupervised task.

3.1.1. *Experiment 2*. Experiment 2 consisted of a non-intercorrelated and an intercorrelated condition. Stimulus items in both conditions depicted imaginary animals that were made up of seven attributes: type of head, body, texture, tail, legs, habitat and time of day pictured. Each attribute could take on one of three values. For example, the time of day could be 'sunrise', 'nighttime' or 'midday'.

Training items in the non-intercorrelated condition preserved only one pairwise correlation between stimulus attributes. All of the stimulus items thus conformed to one the patterns shown in table 1. If the first stimulus dimension encoded the head of an imaginary animal and the second stimulus dimension encoded the body, then knowledge about the type of head an animal possessed would allow prediction of what type of body it had and vice versa. The remaining five dimensions were not correlated so that they were not useful for prediction.

Data items in the intercorrelated condition had six of these pairwise correlations. The first four dimensions of these items were constrained to vary together like the first two dimensions in the non-intercorrelated condition (see table 1). Since four dimensions were involved and because the correlations were interrelated, there were six pairwise correlations in the training items for the intercorrelated condition (e.g. cor(A,B), cor(A,C), cor(A,D), cor(B,C), cor(B, D), cor(C, D), where cor(Y,Z) indicates the values of dimensions Y and Z correlate).

In the learning phase for both conditions, subjects were told that they were participating in a visual memory experiment and viewed the stimulus items for four blocks (four passes through all of the training items). Each item appeared once per block in a random order. The only difference between the two conditions (non-intercorrelated and intercorrelated) was the abstract structure of the items that were used during training.

In the test phase of the experiment, subjects viewed a novel set of 45 stimulus item pairs. Each member of the pair had two obscured attribute values (e.g. the locations

Table 1. The logical structure of the stimulus items for the non-intercorrelated and intercorrelated conditions in Experiment 2 and 3 of Billman and Knutson (1996). The seven columns 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.

Experiment 2			Experiment 3		
Non-intercorrelated condition			Non-intercorrelated Condition		
1 1 x x x x x	2 2 x x x x x	3 3 x x x x x	1 1 1 1 1 1 x	2 2 1 1 1 1 x	3 3 1 1 1 1 x
Intercorrelated condition			1 1 1 1 2 2 x	2 2 1 1 2 2 x	3 3 1 1 2 2 x
1 1 1 x x x x	2 2 2 x x x x	3 3 3 x x x x	1 1 1 1 3 3 x	2 2 1 1 3 3 x	3 3 1 1 3 3 x
			1 1 2 2 1 1 x	2 2 2 2 1 1 x	3 3 2 2 1 1 x
			1 1 2 2 2 2 x	2 2 2 2 2 2 x	3 3 2 2 2 2 x
			1 1 2 2 3 3 x	2 2 2 2 3 3 x	3 3 2 2 3 3 x
			1 1 3 3 1 1 x	2 2 3 3 1 1 x	3 3 3 3 1 1 x
			1 1 3 3 2 2 x	2 2 3 3 2 2 x	3 3 3 3 2 2 x
			1 1 3 3 3 3 x	2 2 3 3 3 3 x	3 3 3 3 3 3 x
			Intercorrelated condition		
			1 1 1 1 x x x	2 2 2 2 x x x	3 3 3 3 x x x

where the tail and head should have been were blacked out) so that in the intercorrelated condition information about only one correlation was available from each test item. The purpose of blocking dimensions was to query learning on only one correlation at a time.

Subjects were asked to evaluate the remaining five attributes that were visible and to choose 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 the correlations present in the items viewed during the study phase and the other was considered 'incorrect' because it did not preserve the correlations.

The basic result from Experiment 2 was that the 'correct' item was chosen more often in the intercorrelated condition than in the non-intercorrelated condition (73% vs. 62%). This finding supports the hypothesis that extracting a category's structure is facilitated by intercorrelated dimensions.

3.1.2. *Experiment 3.* An alternative explanation of the results from Experiment 2 is that a larger number of pairwise correlations in the intercorrelated condition (relative to the non-intercorrelated condition) facilitated learning. To test this explanation, the number of pairwise correlations in the non-intercorrelated and intercorrelated conditions were equated in Experiment 3.

In the non-intercorrelated condition, the items had three isolated pairwise correlations. The abstract structure of the items constrained the first six dimensions into three orthogonal pairs of correlated dimensions.

Items in the intercorrelated condition had three interrelated correlations. The first three dimensions were all correlated which created three pairwise correlations (e.g. cor(A,B), cor(B,C), cor(A,C)). Thus, the number of pairwise correlations in the non-intercorrelated and intercorrelated conditions were equal, but the relationship between these pairs varied between conditions. Example stimulus items for both

conditions had the abstract structure shown in table 1 under the Experiment 3 heading.

Experiment 3 used the same training and test procedure as Experiment 2. The basic result from Experiment 3 confirmed the results of Experiment 2 in that the 'correct' item was chosen more often in the intercorrelated condition than in the non-intercorrelated condition (77% vs. 66%).

3.1.3. *Modelling results*. uSUSTAIN was trained in a manner analogous to how subjects were trained by using four randomly ordered learning blocks. No feedback was provided and all stimulus items were encoded as being members of the same category. In order for uSUSTAIN to mimic the forced choice nature of the test phase, a response probability was calculated for each of the two items. The ultimate response of the network was towards the item in the forced choice that had the strongest response probability.

uSUSTAIN was run numerous times on both conditions in both experiments and the results were averaged (see table 3). The best fitting parameters for both Experiment 2 and 3 (one set of parameters was used to model both studies) are shown in table 2 under the unsupervised column. For both experiments, uSUSTAIN correctly predicts greater accuracy in the intercorrelated condition than in the non-intercorrelated condition (see table 3).

In Experiment 2, uSUSTAIN's most common solution in the non-intercorrelated condition was to partition the studied items into three clusters. The three clusters encoded each of the three possible values of the correlation between the first two dimensions. Accordingly, attention was shifted to the first two stimulus dimensions. For the non-intercorrelated condition in Experiment 3, uSUSTAIN also created three clusters, but the nature of the clusters varied across simulations. In each simulation, uSUSTAIN focused on one of three pairwise correlations and largely ignored the other two. For instance, during training uSUSTAIN might create three

Table 2. uSUSTAIN's best fitting parameters for the studies considered.

function/adjusts	symbol	unsupervised	first/lastname	inf/class	six types
learning rate	η	0.0966	0.0834	0.0796	0.159
cluster competition	β	6.40	7.4922	1.897	1.930
decision consistency	d	1.98	16.480	16.093	6.635
attentional focus	r	10.0	0.4022	6.102	7.156
threshold	τ	0.5	0.3733	0.755	0.701/0.552
distinct focus	$\lambda_{distinct}$	–	3.891	–	–
category focus	λ_{label}	–	–	2.073	–

Table 3. The mean accuracy for humans and uSUSTAIN (shown in parentheses) for Billman and Knutson's (1996) Experiment 2 and 3.

	Non-intercorrelated	Intercorrelated
Experiment 2	0.62 (0.67)	0.73 (0.79)
Experiment 3	0.66 (0.60)	0.77 (0.77)

clusters organized around the third and fourth stimulus dimensions (one cluster for each correlated value pair across the two dimensions) and largely ignore the correlation between the first and second dimensions and the fifth and sixth dimensions. Attention was shifted to the two selected dimensions.

The same dynamics that lead uSUSTAIN to focus on only one correlation in Experiment 3's non-intercorrelated condition led uSUSTAIN to focus on all of the interrelated correlations in the intercorrelated conditions. When uSUSTAIN learns one correlation in the intercorrelated conditions, uSUSTAIN necessarily learns all of the pairwise correlations because the correlated values are all encoded by a common cluster.

uSUSTAIN's operation suggests some novel predictions: (1) Learning about a correlation is more likely to make learning about another correlation more difficult when the correlations are not interrelated; (2) when correlations are interrelated, either all of the correlations are learned or none of the correlations are learned. Both of these predictions have been tested and verified with human subjects (Gureckis and Love, 2002).

3.2. *Modelling sorting behaviour with uSUSTAIN*

In category construction (sorting) studies, human subjects are given cards depicting stimulus items and are instructed to freely sort the cards into piles that naturally order the stimuli. In other words, subjects sort the stimuli into the natural substructures of the category without any supervision. Billman and Knutson's (1996) studies found that subjects preferred stimulus organizations in which the stimulus dimensions were intercorrelated. Interestingly, category construction studies reveal a contrasting pattern—subjects tend to sort stimuli along a single dimension. This behaviour persists despite the fact that alternate organizations exist that respect the intercorrelated nature of the stimuli (Medin *et al.* 1987).

uSUSTAIN was applied to the sorting data from Medin *et al.*'s (1987) Experiment 1 in hopes of reconciling the apparently contradictory findings. In Experiment 1, subjects were instructed to sort stimuli into two equal 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 4. The finding is that all subjects chose to sort the cards along a single dimension as opposed to sorting stimuli according to their intercorrelated structure (i.e. the family resemblance structure shown in table 4).

Table 4. The logical structure of the four perceptual dimensions used in Medin *et al.* (1987). The stimuli sorted in columns according to family resemblance.

1	1	1	1		2	2	2	2
1	1	1	2		2	2	2	1
1	1	2	1		2	2	1	2
1	2	1	1		2	1	2	2
2	1	1	1		1	2	2	2

When uSUSTAIN was applied to the stimulus set from Experiment 1 it was constrained to create only two piles (i.e. clusters) like Medin *et al.*'s subjects. This was accomplished by not allowing uSUSTAIN to recruit a third cluster. This modification proved to be unnecessary as an unrestricted version of uSUSTAIN recruited two clusters in 99% of simulations. uSUSTAIN was presented with the items from table 4 for 10 random training blocks to mirror subjects' examination of the stimulus set and their ruminations as to how to organize the stimuli. To evaluate the performance of the model, we looked at how uSUSTAIN's two clusters were organized. Using the same parameters that were used in the Billman and Knutson (1996) studies listed in table 2, uSUSTAIN correctly predicted that the majority of sorts (99%) will be organized along one stimulus dimension.

uSUSTAIN'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 is directed towards stimulus dimensions that consistently match at the cluster level. This leads to certain dimensions becoming more salient over the course of learning. The dimension that develops the greatest salience over the course of learning becomes the basis for the unidimensional sort. Because of the way attention is updated over the course of trials, uSUSTAIN predicts that which dimension a subject chooses to sort the stimuli on is dependent on the order in which the they encounter the stimuli. Gureckis and Love (2002) recently tested uSUSTAIN's prediction in a sequential sorting study and confirmed that stimulus ordering plays a role in determining which dimension subjects choose as the basis for their sort.

Interestingly, uSUSTAIN was able to account for both Billman and Knutson (1996) data and the Medin *et al.* (1987) data with a single set of parameters despite the differences in the findings. uSUSTAIN's combined account of Billman and Knutson's (1996) studies and Medin *et al.* (1987) suggest that the salience 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. In cases where the total number of correlations is modest and the correlations are weak and not interrelated (as in Medin *et al.* 1987), uSUSTAIN predicts that stimuli will be organized along a single dimension.

4. Modelling supervised learning

The success of uSUSTAIN in the unsupervised learning studies considered above is encouraging, but not surprising. The recruitment procedure that uSUSTAIN employs is a generalization of unsupervised recruitment procedure used by the original SUSTAIN model. Thus, uSUSTAIN can account for any unsupervised learning study that SUSTAIN can. The true test of uSUSTAIN lies in its ability to fit supervised learning data. The following section considers uSUSTAIN's application to studies in supervised learning that SUSTAIN has successfully fit using the error driven recruitment rule. Where appropriate, comparisons between the performance of SUSTAIN and uSUSTAIN will be made.

4.1. *Modelling Medin* et al*'s (1983) comparison of identification and classification learning study*

We begin our evaluation of uSUSTAIN in supervised learning tasks by looking at Medin *et al.*'s 1983 study comparing identification and classification learning. Medin *et al.* (1983) found that under some circumstances identification learning (assigning a unique category label to each stimulus item) is actually easier than partitioning the same stimuli into two categories. Medin *et al.* attributed their surprising result, which existing models cannot account for, to the nature of the stimulus set employed. The stimuli in Medin *et al.*'s study consisted of nine photographs of female faces. Unlike the stimuli used in typical laboratory studies of category learning, these stimuli contain a large amount of idiosyncratic information and are easily discriminated from one another.

Medin *et al.* referred to their identification condition as the first name condition because subjects learned the first name of each person depicted in the photographs, whereas the classification condition was referred to as the last name condition because here subjects learned to assign the photographs to one of two possible 'families'. Table 5 shows the logical structure of the first and last name conditions. In both conditions, subjects were trained using a supervised learning procedure until they correctly classified all nine items for consecutive blocks or until they completed the sixteenth learning block. Feedback was provided after each response.

The results from Medin *et al.* (1983) are shown in table 6. The mean number of learning blocks required by subjects was 7.1 in the first name condition and 9.7 in the last name condition. Overall response accuracy was roughly equivalent in the two condition, even though chance guessing should have favoured the last name

Table 5. 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

Table 6. Human performance and uSUSTAIN's (in parentheses).

Problem type	Blocks required	Proportion reaching criterion	Overall accuracy
First name	7.1 (7.2)	1.00 (1.00)	0.84 (0.85)
Last name	9.7 (10.5)	0.91 (0.95)	0.87 (0.88)

condition (i.e. pure guessing would result in 1/2 correct in the last name condition compared to 1/9 in the first name conditions).

In order to model this data, certain assumptions had to be made about the nature of the input representation. In addition to the experimentally controlled features of hair colour, smile type, hair length, and shirt colour, photographs of human faces contain a large amount of extra information. Because subjects were sensitive to this idiosyncratic information, an additional input dimension was added to each item. The added dimension had the effect of making each stimulus more distinctive. To account for the likely salience differences between this idiosyncratic dimension and the experimentally controlled stimulus dimensions, an additional parameter $\lambda_{distinct}$ was added to uSUSTAIN. The additional parameter allowed uSUSTAIN to initially weight the distinctive dimension differently than the other dimensions (dimensions normally have an initial λ of 1).

uSUSTAIN was able to capture the correct pattern of results with the parameterization shown in table 6 under the heading first/last name. uSUSTAIN correctly predicts that overall accuracy between the two conditions should be roughly equal (despite the fact that chance guessing favors the last name condition), that more learning blocks should be required in the last name condition than in the first name condition, and that a greater proportion of learning runs should reach criterion in the first name condition than in the last name condition.

uSUSTAIN recruited more clusters (nine for each simulation) in the first name condition than in the last name condition (the modal solution involved two clusters). It is important to note that abstraction did not occur in the first name condition because each cluster responded to only one of the nine items, but it did occur in the last name condition. uSUSTAIN's behaviour is driven by the distinctiveness of the stimuli (which was modelled using the added dimension). With distinctive stimuli, 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). Note that the diminished benefit of abstraction negatively impacts performance in the last name condition, but does not affect the first name condition. uSUSTAIN's account of the Medin *et al.* data has been verified by further experimentation with well controlled laboratory stimuli (Love 2000c).

Overall, the Medin *et al.* simulations provide strong support for the unified recruitment rule. The primary difference between the original version of SUSTAIN and uSUSTAIN's solution to this data is that SUSTAIN recruited seven clusters in the last name condition compared to the two clusters recruited by uSUSTAIN. uSUSTAIN's account of the data is actually more in accord with Medin *et al.*'s account which stressed the role of abstraction in the last name condition.

4.2. *Modelling inference and classification learning*
In this section, uSUSTAIN is fit to a series of experiments from Yamauchi and Markman (1998) and Yamauchi *et al.* (2002) comparing human 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. This is in contrast to classification learning in which the value of all perceptual dimensions are known and the category label is being queried. However, just like with classification learning, inference learning is properly characterized as supervised in that the learner receives corrective feedback

on every trial. After receiving feedback the stimulus information available to the learner is equivalent in both inference and classification learning. Despite the similarities, these two learning modes focus human learners on different sources of information and lead to different category representations.

In particular, inference learning tends to focus subjects on the internal structure or prototype of each category whereas classification learning tends to focus subjects on information that discriminates between the two categories. Accordingly, the difficulty of mastering a learning problem can be dependent on which of these two learning modes in engaged.

Yamauchi and Markman (1998) trained subjects using inference and classification on the linear category structure shown in table 7, while Yamauchi *et al.* (2002) trained subjects using the nonlinear category structure that appears in table 8. In both studies, subjects completed 30 blocks of training or until they surpassed 90% accuracy for a three block span. The perceptual dimensions were form, size, colour, and position.

The basic interaction observed between inference and classification learning is that inference 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. The complete pattern of results for these two studies is shown in table 9. The acquisition patterns found support the notion that inference learning focuses subjects on the internal structure of each category whereas classification learning focuses subjects on information that discriminates between the categories.

The procedure used to train uSUSTAIN mimicked the procedure used to train humans. The mean number of blocks required for uSUSTAIN to reach criterion in each condition is shown in table 9. The best fitting parameters are shown in table 2 under the heading inf/class. Note that an additional parameter, λ_{label} (category focus), was utilized in these simulations. The category focus parameter governs how

Table 7. 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

Table 8. 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 9. The mean number of inference and classification learning blocks required for humans and uSUSTAIN (shown in parentheses). Subjects (and simulations) not reaching the learning criterion were scored as a 30 (the maximum number of blocks).

	Inference	Classification
Linear	11.5 (4.6)	13.4 (12.8)
Non-linear	27.4 (29.9)	10.3 (11.3)

Table 10. The modal number of clusters recruited by uSUSTAIN for the inference and classification learning problems.

	Inference	Classification
Linear	2	8
Non-linear	2	6

much attention is placed on the category label at the beginning of a learning episode (akin to a subject's initial biases when entering the laboratory). Given the important organizational role that we hypothesize the category label plays in light of the results from Yamauchi and Markman (2000), we wanted to give uSUSTAIN the option of placing more importance on the category label at the start of training. Indeed, uSUSTAIN differentially weighted the category label relative to the perceptual dimensions which all have an initial tuning of 1.

uSUSTAIN did a good job of capturing the basic pattern of the data (see table 9). The model correctly predicts that inference learning is better suited to linear category structures than it is non-linear category structures. Quantitatively, uSUSTAIN overestimates the strength of this interaction. Table 10 displays the modal number of clusters recruited. In accord with Yamauchi *et al.*'s account of the data, uSUSTAIN suggests that learners focus on the prototype of each category in inference learning, but memorize exemplars in classification learning. The focus on the category prototype was very helpful for the linear category structure, but disastrous for the nonlinear category structure because category prototypes are not sufficient to segregate the category members correctly.

Overall, these simulations provide further support for the unified recruitment rule. The primary difference between the results reported here and those of the original version of SUSTAIN is that uSUSTAIN predicts greater abstraction in inference for non-linear category structures. The original version of SUSTAIN makes numerous prediction errors in this condition which leads to a larger number of clusters being recruited. Although the quantitative fit of SUSTAIN is superior, uSUSTAIN successfully captures the qualitative pattern of the data. Furthermore, uSUSTAIN's

account of the data is actually more in accord with Yamauchi *et al.*'s account than is SUSTAIN's.

4.3. *Shepard* et al.*'s (1961) six classification problems*

The final study that we will consider is Shepard *et al.*'s (1961) classic experiments on human category learning. In this study, human subjects learned to classify eight items that varied on three perceptual binary dimensions (shape, size and colour) into two categories (four items per category). On every trial, subjects assigned a stimulus to a category and feedback was provided. Subjects were trained for 32 blocks or until they completed four consecutive blocks without an error where a block is defined as the presentation of each stimulus item in a random order. Six different assignments of items to categories were tested that varied in difficulty. The logical structure of the six problems is shown in table 11.

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 on the first two dimensions with an irrelevant third 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 *et al.* (1994a) replicated Shepard *et al.* (1961) with more human subjects and traced out learning curves. Figure 3 shows the learning curves for the six problem types. The basic finding is that type I is learned faster than type II which is learned faster than types III–V which are learned faster than type VI. This data is particularly challenging for learning models as most models fail to predict type II easier than types III–V. The only models known to reasonably fit these data are ALCOVE (Kruschke 1992) and RULEX (Nosofsky *et al.* 1994).

The procedure used to simulate uSUSTAIN mimicked the procedure used to collect data from human subjects. uSUSTAIN's best fit of Nosofsky *et al.*'s (1994a)

Table 11. The logical structure of the six classification problems tested in Shepard *et al.* (1961) is shown. The perceptual dimensions (e.g. large, dark, triangle, etc.) were randomly assigned to an input dimension for each subject.

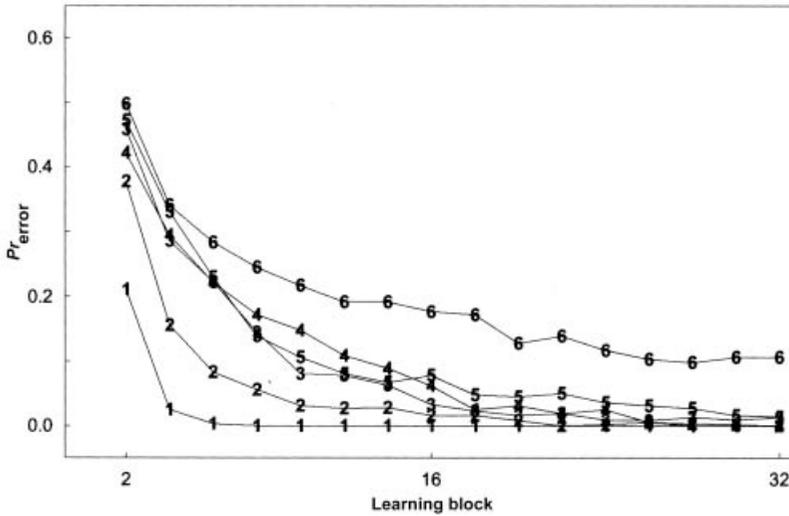| Stimulus			I	II	III	IV	V	VI
1	1	1	A	A	B	B	B	B
1	1	2	A	A	B	B	B	A
1	2	1	A	B	B	B	B	A
1	2	2	A	B	A	A	A	B
2	1	1	B	B	A	B	A	A
2	1	2	B	B	B	A	A	B
2	2	1	B	A	A	A	A	B
2	2	2	B	A	A	A	B	A

Figure 3. Nosofsky *et al.*'s (1994a) replication of Shepard *et al.* (1961).
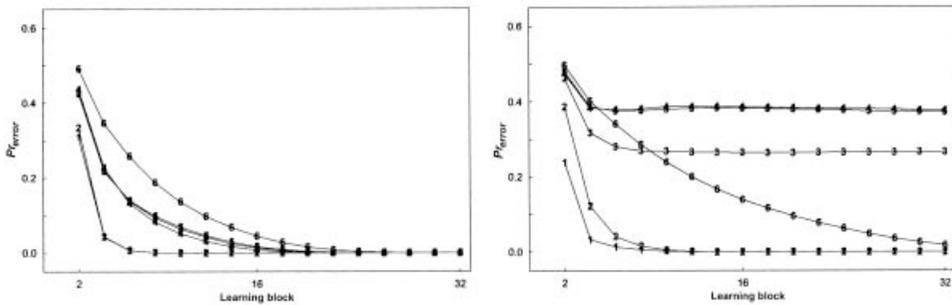

Figure 4. Two different fits of uSUSTAIN's to Nosofsky *et al.*'s (1994a) data is shown averaged over many simulations. The fit on the right used a τ parameter of 0.701 and the fit on the left used a τ of 0.552 (see table 2).

data is shown in the left panel of figure 4. uSUSTAIN does a good job at capturing the data except that it incorrectly predicts that types I and II are of equal difficulty. Only the first two stimulus dimensions of the type II problem are relevant to classification and this problem is naturally captured by four clusters, whereas only one dimension is relevant for the type I category and it is naturally captured by two clusters (one for each value of this dimension). uSUSTAIN's modal solution is to recruit four clusters for both the type I and type II problems which leads to its incorrect prediction. One solution to this misprediction for the type I and type II problems is to lower the τ recruitment parameter so that uSUSTAIN will tend to recruit fewer clusters. Doing so results in two clusters being recruited for the type I problem and four for the type II problem, and leads to the correct ordering of these two problems (see the left panel of figure 4). Unfortunately, a side effect of adjusting the recruitment parameter is that the rule-plus-exception categories (types III–V) no longer converge to 100% accuracy. Early in learning, the correct order for all six problems is displayed. However, in types III–V certain exception items are more

similar to a cluster belonging to the opposing category than they are to any cluster belonging to the appropriate category. Because of this, uSUSTAIN cannot create a new cluster to encode these exceptions and continually makes prediction errors for these items.

uSUSTAIN's fit of the Shepard *et al.* problems was inferior to the fit obtained using the original version of SUSTAIN. However, uSUSTAIN can capture many aspects of the data. The qualitative fit of the simulations shown in the right panel of figure 4 captures quite a bit of the data (though the predicted difference between types I and II is not observed). The simulations shown in the right panel compliment those found on the left and predict the correct ordering of problem difficulty in the early blocks of learning. However exceptions are never memorized due to the low recruitment parameter, τ, and learning never converges to 100% accuracy.

The Shepard *et al.* problems represent the most difficult test for the unified recruitment rule as these learning problems stress hypothesis testing and the encoding of exceptions due to *prediction* errors. At this time, it is unclear whether these results indicate fundamental limitations of the recruitment rule or if more work is needed to improve upon our basic approach.

5. General discussion

SUSTAIN has extended the range of induction tasks that can be modeled by category learning models. The purpose of this paper was to consider a new version of SUSTAIN that unifies SUSTAIN's supervised and unsupervised recruitment procedures. Both of the recruitment procedures used in the original version of SUSTAIN were based on a notion of surprise. The new version of SUSTAIN, referred to as uSUSTAIN, generalized SUSTAIN's unsupervised recruitment rule so that it could also be applied to supervised learning situations. In addition to being a more parsimonious account, uSUSTAIN also makes interesting claims about the relationship between unsupervised and supervised learning. The success of uSUS-TAIN suggests that these two learning procedures differ quantitatively, not qualitatively. In both supervised and unsupervised learning, new clusters are recruited when an item is not sufficiently similar to any cluster of the appropriate category. In the case of unsupervised learning, all items belong to the same global category, whereas in supervised learning there are multiple contrastive categories.

This approach differs from previous error driven schemes. Despite its simplicity, the unified recruitment procedure has shown to be remarkably successful. In addition to capturing unsupervised learning data, uSUSTAIN can account for many aspects of supervised data sets. In the case of the Medin *et al.* (1983) studies and the inference and classification learning studies, uSUSTAIN's account may be more theoretically well grounded than the one provided by the original version of SUSTAIN.

The one study in which uSUSTAIN did not excel was the Shepard *et al.* problems. This may represent a limiting case for uSUSTAIN as the problems stress hypothesis testing and error driven storage of exceptions. Nevertheless, uSUSTAIN was able to capture some of the key aspects of this data set. Future work will determine whether such supervised data sets can be captured by a generalized unsupervised recruitment procedure or if such studies are outside the privy of the approach. The results presented here are quite promising and suggest that unsupervised and supervised learning are much more alike than they are different.

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

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

Berry, D. C., and Dienes, Z., 1993, *Implicit Learning: Theoretical and empirical issues* (Hillsdale, NJ: Erlbaum).

Billman, D., and Heit, E., 1988, Observational learning from internal feedback: a simulation of an adaptive learning method. *Cognitive Science*, **12**(4): 587–625.

Billman, D., and Knutson, J., 1996, Unsupervised concept learning and value systematicity: a complex whole aids learning the parts. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **22**(2): 458–475.

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

Clapper, J. P., and Bower, G. H., 1994, Category invention in unsupervised learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **20**: 443–460.

Cleermans, A., 1993, *Mechanisms of implicit learning: Connectionist models of sequence processing* (Cambridge, MA: MIT).

Gureckis, T., and Love, B. C., 2002, Who says models can only do what you tell them? unsupervised category learning data, fits, and predictions. In *Proceedings of the 24th Annual Conference of the Cognitive Science Society*, Mahwah, NJ, Lawrence Erlbaum Associates, in press.

Hayes, N., and Broadbent, D. E., 1988, Two modes of learning for interactive tasks. *Cognition*, **28**: 249–276.

Hock, H. S., Malcus, L., and Hasher, L., 1986, Frequency discrimination: assessing global and elemental letter units in memory. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **12**: 232–240.

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

Lewicki, P., 1986, *Nonconscious social information processing* (New York: Academic Press).

Love, B. C., 2002a, Comparing supervised and unsupervised category learning. *Psychonomic Bulletin and Review*, in press.

Love, B. C., 2002b, The multifaceted nature of unsupervised category learning. *Psychonomic Bulletin and Review*, in press.

Love, B. C., 2000c, Learning at different levels of abstraction. *Proceedings of the Cognitive Science Society*, pp. 800–805.

Love, B. C., Markman, A. B., and Yamauchi, T., 2000, Modeling classification and inference learning. *Proceedings of the Fifteenth National Conference on Artifical Intelligence*, pp. 136–141.

Love, B. C., and Medin, D. L., 1998a, Modeling item and category learning. In *Proceedings of the 20th Annual Conference of the Cognitive Science Society* (Mahwah, NJ: Lawrence Erlbaum Associates), pp. 639–644.

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

Love, B. C., Medin, D. L., and Gureckis, T., 2002, SUSTAIN: a network model of human category learning. *Psychological Review*, forthcoming.

Luce, R. D., 1959, *Individual choice behavior: a theoretical analysis* (Westport, CO: Greenwood).

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

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

Nosofsky, R. M., Gluck, M. A., Palmeri, T. J., McKinley, S. C., and Glauthier, P., 1994a, 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., and McKinley, S. C., 1994b, Rule-plus-exception model of classification learning. *Psychological Review*, **101**(1): 53–79.

Ratcliff, R., 1990, Connectionist models of recognition memory: constraints imposed by learning and forgetting functions. *Psychological Review*, **97**: 285–308.

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

Shepard, R. N., Hovland, C. L., and 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 & Reasoning*, **3**: 81–110.

Widrow, B., and Hoff, M. E., 1960, Adaptive switching circuits. In *IRE WESCON Convention Record* (New York), pp. 96–104.

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

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

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