Chapter 6

Integrating Associative Models of Supervised and Unsupervised Categorization

Jan Zwickel and A. J. Wills

Categorization—dividing the world into groups of things—is one of the core mechanisms behind many cognitive abilities. Categorization enables people to cope with the otherwise overwhelming complexity of objects and situations by reducing information. It also allows people to generate predictions by generalizing what they already know to novel situations. For example, if you meet, say, a new doctor, you can make certain inferences about how they will act based on what you already know about the category "doctor." As two situations are rarely (if ever) identical such an ability seems essential to everyday life. On the other hand, overgeneralization can be problematic. For example, it might be useful to have a "cat" category that covers both your own pet and other similar animals. It would, however, be potentially disastrous if you generalized what you knew about this category to the first lion you met.

Most of the psychological research into our ability to categorize has employed a *supervised learning* paradigm (e.g., Bruner, Goodnow, & Austin, 1956; Gluck & Bower, 1988; Medin & Schaffer, 1978). Prototypically, supervised learning assumes the presence of a perfect teacher, who observes *every* category choice and provides correct feedback on *every* decision (i.e., which category was the correct one to choose). This technique has undoubtedly revealed important information about the psychological processes of categorization. Nevertheless, it seems unlikely that this level of information is commonly available outside of the laboratory.

Free classification, which can also be described as *unsupervised learning,* is a methodology that goes to the opposite extreme. Participants are typically asked to partition stimuli into groups that seem reasonable or "natural" to them (e.g., Ahn & Medin, 1992; Bersted, Brown, & Evans, 1969; Medin, Wattenmaker, & Hampson, 1987; Wills & McLaren, 1998). No feedback on categorization decisions is given.

In this chapter, we describe three basic associative theories of learning as they might be applied to the problem of category learning. First, we describe *Hebbian learning*, which was one of the earliest formal associative theories of learning. Second, we describe *competitive learning*, which can, in some ways, been seen as a development of Hebbian learning, and is a theory of *unsupervised learning*. Next, we describe the *Rescorla–Wagner* theory, which is a theory of *supervised learning*. We then argue that there is a need for a mechanism that can use feedback when it is available (supervised learning) but will continue to learn if feedback is absent (unsupervised learning). We propose a possible mechanism that involves adding certain aspects of the Rescorla–Wagner theory to competitive learning. Our proposed system makes a clear prediction about people's behavior in a free-classification task, and we describe how we have begun to test that prediction. The chapter ends with a consideration of other theoretical approaches to the problem of category learning, both within the domain of associative learning and more generally.

HEBBIAN LEARNING AND AN INTRODUCTION TO CONNECTIONISM

Hebb (1949) postulated that "when an axon of cell A is near enough to excite a cell B and repeatedly or persistently takes part in firing it, some growth process or metabolic change takes place in one or both cells such that A's efficiency, as one of the cells firing B, is increased" (p. 62). One way of expressing *Hebbian* learning formally is shown in Equation 6.1:

$$\Delta w_{ij} = G a_i a_j \tag{6.1}$$

where a_i and a_j are the activities of two different neurons, i and j, Δw_{ij} is the change in strength of the connection w_{ij} between neuron i and neuron j, and G is a parameter that affects the rate of learning. For simplicity, let's assume that a_i and a_j can take only values of 0 and 1. If a_i and a_j are both 1, then the connection strength increases by G. However, if either i or j is not active, the equation is zero and the connection strength doesn't change. By this equation, strong connections are developed between neurons that are active at the same time and weak connections between neurons that seldom fire together.

Hebbian learning has the potential to acquire the sort of information we need to learn categories. For example, imagine a situation where neuron i represents a feature of a category, and neuron j represents the category label. If feature i is typical of category j then a strong link will form between the two, and so a novel object that has feature i is likely (other things being equal) to be categorized as a member of group j. However, it's worth making clear at this point that we are not assuming there is necessarily a *single* neuron that uniquely represents, say, the category "football player." This may be the case, or it might be that categories are represented as a pattern of activity distributed over many neurons (see Page, 2000, for an excellent review of this issue).

In what follows, the neuron or neurons that represent a stimulus feature are called an *input unit* and the neuron or neurons that represent a category are called an *output unit*. The strength of the connection between an input unit and an output unit represents the strength of the learned association between a stimulus feature and a category. The set of input units, output units, and the connections between them is described as a connectionist, or neural, network.

COMPETITIVE LEARNING

This section is an introduction to competitive networks as a model of unsupervised category learning. In an unsupervised category-learning situation, stimuli are presented but the category membership of those stimuli is not given. The network must therefore decide which category the stimulus belongs to. One way of doing this is to assume that the output unit made most active by the presentation of the stimulus is the one that represents the category.

In a standard competitive network (and many other kinds of networks), the activity of an output unit is the sum of the activities of the input units multiplied by the connection weights from the inputs units to the output unit. Formally,

$$o_j = \sum_i a_i w_{ij} \tag{6.2}$$

where o_j is the activity of output unit j, a_i is the activation of input unit i, and w_{ij} is the connection strength between input unit i and the output unit j.

Assuming that connection strengths start with random values, it is likely that the first stimulus presented will make one of the output units more active than any of the others. Under competitive learning, this output unit is considered the "winner" and its activity is set to one. The activity of all other output units is set to zero. A learning rule can then be applied. For example, if one applies Hebbian learning, the connection strengths between the active input units and the winning output units would be increased. All other connection strengths would remain unchanged.

Through the application of a learning rule such as Hebbian learning, the winning output unit comes to more strongly represent the presented stimulus, and hence is more likely to win again if a similar stimulus is presented in the future (similar in the sense of having many features in common). However, if a substantially different stimulus is presented, it is possible that a different output unit will win. In turn, that output unit will begin, through the action of the learning rule, to more strongly represent this substantially different stimulus and those similar to it. The system has therefore begun to develop the ability to categorize, despite the absence of external information about category membership.

One problem with using basic Hebbian learning in this way is that, in practice, one output unit can often end up representing all the stimuli. In an attempt to reduce this problem, Rumelhart and Zipser (1986) introduced further competition between the output units by introducing a limit on the sum of all connection strengths to any given output unit. This means that, after a certain point, the only way a connection strength to a particular output unit can increase is if other connection strengths to that output unit decrease. This encourages output units to specialize on different kinds of feature patterns.

Rumelhart and Zipser restrict the maximum level of connection strengths by assuming the presence of a kind of "decay" process that operates every time the connection strengths are changed. Formally, if unit j loses,

$$\Delta w_{ij} = 0 \tag{6.3a}$$

and hence connection strengths to losing output units remains unchanged but, if unit j wins then

$$\Delta w_{ij} = G\frac{a_i}{n} - Gw_{ij} \tag{6.3b}$$

where G is again a learning parameter, a_i is one if input unit i is active but zero otherwise, and n is the number of active input units. Note that the inclusion of n means that the increase in connection strength is greater the fewer the number of active input units. This compensates an output unit that is supported by few input neurons.

Equations 6.3a and 6.3b implement a mechanism similar but not identical to Hebbian learning. If input unit i is active when the winning unit is active, the connection strengths increase. However, the connection strengths also decay a bit because of the last term of Equation 6.3b. Inclusion of the connection strength w_{ij} in that decay process means that strong connections decay more rapidly than weak connections. If input unit i is not active, then there is no increase in connection strength. However, the decay process ($-Gw_{ij}$) still operates. One outcome of the learning rules expressed in Equations 6.3a and 6.3b is that, once the sum of connection strengths to a given output node reaches 1 it stays there—further learning can only redistribute the total weight across different connections.

Competitive networks suffer from a fundamental limitation. If the task is learning to partition patterns into groups on the basis of overall similarity (measured in terms of feature overlap), then competitive learning may succeed. However, if the task requires overall similarity to be ignored and category responses to be made on the basis of a particular subset of features, then competitive learning will necessarily fail. For example, imagine that you have three objects in front of you to put into two groups—a newspaper, a beer bottle, and a plastic bottle. In terms of overall similarity, the two bottles seem likely to form one category and the newspaper another. However, if the task requires you to separate the items into "recyclable" and "nonrecyclable" then (in some districts), the newspaper and the glass bottle go into the recycling bin, but the plastic bottle goes into the landfill bill.

How could a neural network create a different grouping than one that is based on overall similarity? One possible solution is feedback. Feedback provides information from outside, enriches gathered information, and thereby changes the priority of different features through the knowledge of an "expert." This external knowledge is not commonly used in a competitive network; it is the realm of another class of networks that engage in *supervised* learning. In the next section, we consider one such theory.

THE RESCORLA–WAGNER MODEL (OR DELTA RULE)

The Rescorla–Wagner model (Rescorla & Wagner, 1972) was developed in the domain of animal-learning theory, although similar models can be traced back to Widrow and Hoff's (1960) work on electronic switching circuits. In connectionist modeling, the model is often referred to as the delta rule.

Roughly speaking, the model works in the following way. After a stimulus is presented, the model predicts whether or not an outcome will occur. The environment then provides the model with feedback about whether the predicted outcome did in fact occur. If the model made the correct prediction then it assumes there is no need for further learning. However, if there is a discrepancy between the model's prediction and the feedback, learning takes place by adjusting connection strengths. These are adjusted in a manner that should reduce the error in future. How does the model achieve this? In what follows we describe the model in terms as similar as possible to those used in our description of competitive learning. As a result, the terminology is more similar to that used in connectionist modeling than that originally used by Rescorla and Wagner. The basic concepts, however, remain unchanged.

The model has, as before, a set of input units and a set of output units and associative connections from the former to the latter. As with competitive networks, the activity of an output unit is the sum of the activities of input units multiplied by their corresponding connection strengths. The learning rule works in the following way. If output unit j should (on the basis of feedback) be highly active, but the activation coming from the input units is insufficient, the connection strengths from active input units increase. On the other hand, if output unit j should not be active, but activation is coming from the input units, then the connection strengths between active input units and output unit j is reduced. Formally, the change in connection strength between input unit i and output unit j is

$$\Delta w_{ij} = G\left(\lambda_j - \sum_k a_k w_{kj}\right)a_i \tag{6.4}$$

where a_i represents the activity of input unit i, λ_j represents the correct activation of the output unit j, provided by the feedback from the environment, and k sums over all input units. The term in parentheses denotes the difference between the model's prediction and the feedback. This difference is weighted by the activation of the input unit—only active input units can drive changes in connection strength in this system. Stone (1986) demonstrates that the delta rule essentially carries out the equivalent of multiple linear regression.

The delta rule is at the heart of many models of supervised categorization (e.g., Gluck & Bower, 1988; Kruschke, 1996; McClelland & Rumelhart, 1985). In this section, we outline one simple way in which it can be used, based closely on Gluck and Bower's work.

In the system we consider, features of objects are represented by input units and categories are represented by output units. Input units have an activity of one if the feature they represent is present, and zero otherwise. The output unit representing the correct category is assumed to have a λ of one, whereas all other category units have a λ of zero.

The system as described has two well-known properties (see, e.g., Minsky & Papert, 1969). First, if there is a configuration of connection strengths that yields all the right answers, the delta rule is guaranteed to find it. Second, there are a number of problems that the delta rule, in this form, is unable to learn. For example, it cannot learn the exclusive-or (XOR) problem. In the XOR problem, there are two input units and one output unit. If just one of those input units is active, then the output unit should be active. However, if both input units are active or inactive, then the output unit should be inactive. For example, when playing a simple card game, you may be able "stick" or "twist," but you have to do one of the two and you can't do both. If "stick" was represented by one input unit and "twist" by another, the delta rule could never learn which of the four possible responses (stick, twist, stick & twist, neither) were allowed. This is because it needs to form a positive connection between "twist" and "allowed," and also between "stick" and "allowed," so stick and twist together would inevitably result in the response "allowed."

The XOR problem can be solved by introducing a different coding scheme; more specifically one can introduce input units that represent the combination of features. These kind of models are commonly referred to as configural cue or unique cue models (Gluck, 1991; Rescorla, 1973). In our card game example, this would mean a third input unit that represents "stick and twist" that is only active when both the "stick" and the "twist" input units are active. This "stick and twist" unit could then form a strong negative connection to the "allowed" output unit. Together with weaker positive connections from "stick" to "allowed" and "twist" to "allowed," the system could solve this XOR problem. This solution effectively involves turning the XOR problem into a different problem that the network can solve.

Another solution to the XOR problem is to introduce a layer of units between the input and output units. These are generally described as "hidden" units and they allow the network to recode the input it receives. For example, it is possible to create a hidden unit that is active only when both input units are active. If this hidden unit has a sufficiently strong *negative* connection to the output unit, then the output unit's activity will be close to zero when both input units are active (the input units will increase the activity of the output unit, but this will be offset by the reduction in activity caused by the hidden unit).

The delta rule needs to be modified before one can apply it to a system with hidden units. This is because the environment provides no direct information about what the "correct" activity of a hidden unit should be, so the error ($\lambda - \Sigma aw$) for a hidden unit does not have an obvious value. One solution is to calculate the error for output units as normal and then pass that error to the hidden units via the connections between the two. This solution is often described as "back-propagation" and is described in detail by Rumelhart, Hinton and Williams (1986). The history of this back-propagation algorithm can be traced back to Werbos (1974).

The back-propagation system has a number of well-known properties, three of which we consider here because they highlight how the system differs from a simpler delta-rule system with no hidden layer. First, Hornik, Stinchcombe, and White (1989) demonstrated that a delta-rule network with a hidden layer is a universal function approximator. Roughly speaking, this means that if there is a stable relationship between the input patterns and the output patterns, then there is a hidden-layer network with a particular pattern of connection strengths that can reproduce it. This contrasts starkly with the simple delta-rule system, which has clear limitations to the patterns it can reproduce.

A second well-known property of back-propagation is that, although there is always a pattern of connection strengths in some hidden-layer network that will reproduce any given function, back-propagation is not guaranteed to find it. Instead it may, during the course of learning, get stuck in what are described as *local minima*. Roughly speaking, this is where the network finds itself in a position where its current performance is not correct but where any small change in connection strengths makes its performance worse than it already is. This again puts back-propagation into stark contrast with the simple delta-rule system, which will always find the solution if one exists.

Back-propagation is also generally considered to be a neurally implausible system. In other words, given what we currently know about the brain, it seems unlikely that neurons could engage in the sorts of processes back-propagation requires. Again, this is in stark contrast to the simple delta-rule system, which can be constructed from the known properties of neurons (see McLaren, 1989).

INTEGRATING SUPERVISED AND UNSUPERVISED LEARNING

So far, we have discussed one model of learning in unsupervised situations (competitive learning) and a different model of learning in supervised situations (the Rescorla–Wagner model). In this section, we consider the merits of integrated theories that can learn in both supervised and unsupervised situations.

Why is the integration of supervised and unsupervised learning desirable? This is most easily illustrated by considering the alternative, which is that the learner must determine in advance whether to engage in supervised or unsupervised learning. If the world was neatly and predictably divided into situations where no feedback is ever received and situations where feedback is always received, then this might not be too much of a problem. However, it seems likely that for most situations feedback is received sporadically and not entirely predictably. One would therefore be likely to encounter the joint problems of ignoring available feedback and failing to learn anything when feedback is unexpectedly absent. There are also potential issues of how information gained by the supervised and unsupervised systems would be integrated.

Another reason why the integration of supervised and unsupervised learning might be desirable is that it reduces the number of theories needed to explain learning. Following the reasoning of Occam (a medieval monk), scientists often argue that if you have a choice between two explanations, both of which explain the available data, you should pick the simpler ("Occam's razor"). It's our contention that the integrated theory we discuss next is a simpler explanation than one that posits separate systems for supervised and unsupervised learning.

AN INTEGRATED MODEL

In the following, we show one way in which a simple associative model of supervised learning (a single-layer delta-rule network) can be integrated with a simple associative model of unsupervised learning (a Rumelhart and Zipser competitive network). Our approach was to start with the delta-rule system and consider how it could be modified to also account for situations where feedback is missing. When feedback is present, the delta rule tries to minimize the difference between the feedback (λ) and the prediction delivered by the weights (Σaw). One way to generalize this principle to situations where feedback is absent is for the network to produce its own feedback signal, which we designate as λ'. If feedback is present, λ' is determined by that feedback. However, when feedback is absent λ' is set to one for the most active output unit and zero for all other output units (an idea borrowed from competitive learning). Such a system does not need to know in advance whether feedback is going to occur, and could be implemented by introducing fixed inhibitory links between the output units (see Wills, Reimers, Stewart, Suret, & McLaren, 2000, for an example of this type of decision mechanism).

Allowing the delta-rule system to generate its own feedback in this way provides a potential integrated model of supervised and unsupervised learning. However, from our earlier discussion of competitive learning it seems likely that such a system would suffer from a potentially serious problem. Like a competitive system with Hebbian learning, there is a real danger that in unsupervised situations one output unit could come to represent all presented stimuli. This is because there are only very rarely situations where stimuli from different categories have absolutely nothing in common. If stimuli from different categories have some common features, then the "winner" of the first stimulus has an advantage when the second stimulus is applied, due to its stronger connections to the features the first and second stimuli have in common.

Rumelhart and Zipser (1986) included further sources of competition besides the "winner-take-all" competition in an attempt to reduce this problem, and we modified the delta rule in a similar way. Specifically, we modified the delta rule so that it included a sort of weight decay process and a process that scales weight changes by the number of active input units. Hence, our modified delta rule is

$$\Delta w_{ij} = G\left(\lambda'_j - \sum_k a_k w_{kj}\right)\frac{a_i}{n} - G\left(\lambda'_j - \sum_k a_i w_{kj}\right)w_{ij} \qquad (6.5)$$

where G is a learning rate parameter, n is the number of active input units, a_i is the activity of input unit i, w_{ij} is the connection strength from input unit i to winning output unit j, and λ' is the internally generated feedback signal discussed earlier. k sums over all input units.

All that we've done here is add the error-correcting component of the delta rule (λ' – Σ*aw*) to Rumelhart and Zipser's learning algorithm (Equation 6.3b). As a result, this integrated model makes predictions about unsupervised learning that differ from those made by Rumelhart and Zipser's competitive learning system. Specifically, our system adds the constraint that connection strengths will change only if the internal feedback signal λ' is not fully predicted by the network (i.e., if λ' – Σ*aw* does not equal zero). This leads to the prediction that effects such as *blocking* (described in the next section) should be observable in unsupervised learning. In contrast, Rumelhart and Zipser's model predicts that blocking will not be observed in unsupervised learning because their learning rule is a variant of Hebbian learning and hence contains no error-correcting component. In later sections, we report data that indicates blocking does occur in free classification, and discuss simulations that show that our system can predict blocking in free classification but that the Rumelhart and Zipser system cannot. In the next section we provide a very brief outline of the phenomenon of blocking for the benefit of those who are not familiar with it.

BLOCKING

The term *blocking* was coined by Kamin (1969) to describe a phenomenon he observed in rats. In Kamin's experiment, rats learned in an initial phase that pressing a bar leads to a food pellet. After this contingency was established, the first phase started. In the first phase, a noise was sometimes presented. When the noise was on, the rat received an electric shock. Rats quickly suppressed responding in the presence of the noise, as measured by frequency of pressing the bar. In the second phase, the noise was always accompanied by a light. When the noise and the light were on, the rats got shocked. Again, the rats avoided pressing the bar during presentations of the noise–light compound. In the final phase, the light was presented alone, and Kamin found that the rats did not suppress responding in the presence of the light. In other words, the rats appeared not to have learned the association between light and shock.

Kamin also ran a control group of rats. The control group was identical to the experimental group with one exception–the control group skipped the first phase (noise only). In the final phase, the control group showed strong suppression in response to the light, thereby indicating that they had learned the connection between the light and the shock in the second phase. Table 6.1 summarizes the design of Kamin's experiment.

How might this difference between the control and experimental groups be explained? Kamin's explanation employed the notion of "surprise." The experimental group had learned that the noise predicts shock in the first phase. Therefore, the shock was not surprising in the second phase. According to Kamin, learning only occurs if the outcome is surprising, so the rats didn't learn the connection between the light and the shock. However, the control group skipped the first phase and therefore was surprised in the second phase by the shock. Consequentially in the second phase, the rats learned the connection between the light and the shock as well.

TABLE 6.1
Kamin's (1969) Blocking Experiment

Group	*Stage One*	*Stage Two*	*Test*
Experimental	N → Shock	LN → Shock	L
Control		LN → Shock	L

Note. "L" denotes a light, "N" a noise.

The delta rule predicts that blocking should occur. Consider a simple representation of the problem where one input unit represents "tone," another "light," and an output unit represents "shock." There are connections from the input units to the output units, which start at zero. Assuming sufficient training, Phase 1 leads to the connection strength between the "tone" input unit and the "shock" output unit being close to λ. Therefore, in Phase 2 when both the "tone" and "light" input units are active and a shock occurs, there is only a very small error ($\lambda - \Sigma aw$) at the "shock" output unit because Σaw is close to λ. This means that the connection strength between the "light" and "shock" units cannot increase substantially, leading to blocking. However, if the first stage is skipped the initial error when the tone and light are presented together is high, so substantial connections form from both input units to the output unit.

Evidence for blocking can also be found in humans. For example, Dickinson, Shanks, and Evenden (1984) demonstrated a blocking effect in the context of a simple computer game. The game involved tanks driving through an invisible minefield. In the experimental condition, participants first experienced an "observation" phase, where they were asked to observe a number of occasions of a tank driving through the minefield and either blowing up or not. Following this, the participants were given the opportunity to shoot at the tanks. Finally, they were asked to rate the effectiveness of the gun in destroying the tanks. In the control condition, the initial observation phase was omitted. Participants in the experimental condition gave lower ratings of the gun's effectiveness than participants in the control condition did. For the experimental group, the development of a "minefield" → "tank explodes" association in the observation phase blocks the development of a "gun fired" → "tank explodes" association in the second phase.

The demonstration of blocking in humans is not limited to this kind of "ratings" task. For example, Martin and Levy (1991) demonstrated blocking in human eyelid conditioning.

EXPERIMENT

The purpose of our experiment was to demonstrate an effect analogous to blocking in the absence of feedback. Previous research demonstrates that category learning can proceed successfully in the absence of feedback (e.g., Homa & Cultice, 1984; Wills & McLaren, 1998). Additionally, the current study follows on from Zwickel and Wills's (2002) demonstration of a blocking-like effect in a situation where some feedback was present, but it was very sparse and not item specific.

The design of the current experiment is shown in Table 6.2; the letters indicate sets of features that make up the abstract stimuli we presented. In Phase 1, examples of Category 1 were created from a base pattern that contained feature sets A and B. Examples of Category 2 were created from a base pattern that contained feature sets C and D.

TABLE 6.2
Experimental Design

	Phase 1	*Phase 2*	*Test*
Category 1	AB	A<u>E</u>	
			<u>E</u>F
Category 2	CD	GF	

Note. Each letter represents a set of six features. For example, Category 2 in Phase 2 contains feature sets G and F. The redundant feature set E is underlined.

Note that the labels Category 1 and Category 2 are essentially arbitrary in a free-classification task—they could be reversed without changing anything in the design or execution of the experiment. As Table 6.2 illustrates, once the participant had mastered the AB versus CD categorization they were transferred to a second categorization. The testing phase started after the participants had mastered this second categorization. The datum of central importance in this design is the category to which the test stimulus presented in the test phase is allocated. The response to a single stimulus is chosen as the dependent variable because subsequent decisions may be contaminated by learning on previous test trials.

Note that feature set E occurs only in situations where the information it provides is redundant. In Phase 2 the stimuli can be identified as Category 1 on the basis of whether they contain "A" features—an association already learned in Phase 1. Hence, through analogy to selective learning effects in tasks with feedback, one might consider that E develops little control over responding. In contrast, G and F may develop more control over responding as they are the only features in Phase 2 that predict the presence of a Category 2 stimulus. If blocking occurs in free classification, one would therefore expect participants to place stimulus EF into Category 2 (i.e., the same category as they used for stimulus GF). This is because F's association to Category 2 is predicted to be greater than E's association to Category 1, due to E, but not F, being blocked.

Method

Participants and Apparatus. Thirty-two psychology students from the University of Heidelberg participated to fulfill partial course requirements or for a small reward. Participants were tested in groups in a quiet computer room. Stimulus presentation was on color monitors connected to standard PCs running the DMDX software package (Forster & Forster, 2003). Responses were collected via the left and right CTRL keys on standard PC keyboards.

Stimuli. Each stimulus was made up of 12 small pictures (hereafter "elements") taken from a set of 72 that have been used in a number of previous experiments (e.g., Jones, Wills, & McLaren, 1998). See Fig. 6.1 for an example stimulus. For any given stimulus, the 12 elements were randomly arranged in a square of three rows with 4 elements in each row, and were surrounded by a gray rectangle outline 5 cm in height and 4 cm in width. Each of the letters A to G in Table 6.2 represents a set of six elements. The stimuli actually presented to participants were generated by random distortion of the base patterns described in Table 6.2. Each element in a base pattern was given a 10% chance of being replaced by a randomly selected element from the other base pattern. An example may be helpful. To create an AB stimulus in Phase 1, the six A elements and the six B elements were randomly arranged in the four-by-three grid of the stimulus. Each element was then given a 10% chance of being replaced by a randomly selected element from set C or D. This method of stimulus construction produces training examples that are composed predominantly of elements characteristic of a particular category but that also exhibit considerable variability.

In order to control for effects of the differential salience of the elements, participants were divided into pairs. The assignment of picture elements

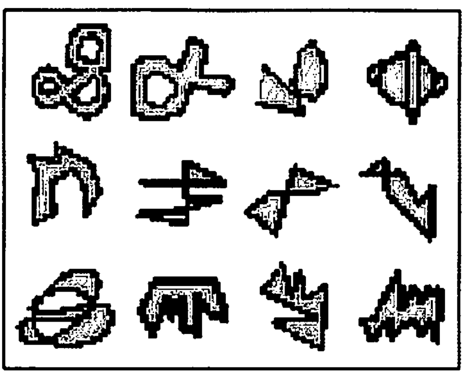

FIG. 6.1. An example stimulus.

to letters was randomly determined for each pair of participants. One participant in each pair received the stimuli described in Table 6.2 whereas the other received a design where E was transposed with F and A with G. Hence, the putatively redundant elements were E for one member of the participant pair, whereas they were F for the other member.

Procedure. The main experiment was preceded by some general written instructions and a brief practice phase to familiarize participants with the procedure. The experiment then proceeded in blocks of 24 trials. On each trial, a stimulus was presented for 800 ms and followed by a mid-gray mask that was presented for 1,200 ms. If a response was not detected within 2,000 ms of stimulus onset, the trial terminated with a message stating that the participant had responded too slowly and asked them to speed up. The participant was then moved on to the next trial.

Each block comprised the sequential presentation of 24 stimuli, 12 from each of the two categories. At the end of each block the percentage of correct responses made by the participant was calculated, but *not* presented to the participant. Clearly, percentage correct has a slightly different interpretation in a free classification task to a task with trial-specific feedback as the relationship between Categories 1 and 2 and the two response keys is arbitrary. Hence, percentage correct was computed by assuming for each block that the key that was pressed most often when stimuli of Category 1 were presented represented the correct answer to Category 1. The other key was assumed to be the correct answer for Category 2. When this "percentage correct" score exceeded 83%, the participant was moved on to the next phase of the experiment. Participants were also moved on to the next phase if they completed six blocks without reaching this criterion.

Results and Discussion

Participants completed a mean of 5.19 blocks in Phase 1, and a mean of 5.25 blocks in Phase 2. Most people did not achieve high levels of performance in this task, with only 4 of the 32 people tested reaching the 83% criterion in both phases. Given the total lack of feedback and the relatively small amount of exposure to complex and unfamiliar stimuli, this is perhaps unsurprising. We therefore decided to exclude only those participants whose performance in Phase 2 was so poor that their response to the test stimulus EF could not be interpreted. In order to be included, a participant's dominant response to Category 1 stimuli in Phase 2 had to be different from their dominant response to Category 2 stimuli in that phase. For example, if when a Category 1 stimulus was presented the participant was more likely to press the right-hand key, then in order to be included they had to be more likely to press the left-hand key in response to Category 2. In other words, they should not predominantly assign the stimuli of both categories to the same key.

Sixteen of the 32 participants passed this criterion. However, one would not expect participants to demonstrate blocking unless they have learned the Phase 1 categorization to some extent. We therefore divided our remaining participants into "learners" and "nonlearners" by applying the same criterion to performance in Phase 1. Finally, we classified all 16 participants on the basis of whether their response to the EF test stimulus was consistent or inconsistent with blocking. Their response was classified as consistent if they responded to EF using the key they had predominantly used to respond to GF in Phase 2. If they used the other key, they were classified as inconsistent with the blocking hypothesis.

Looking at Table 6.3, one can see that the majority of the learner's classifications of the test stimulus were consistent with the blocking hypothesis. Tested against a null hypothesis of random responding, the evidence for our blocking hypothesis misses significance by the narrowest of margins, $p = 0.05$, one-tailed, on a binomial test. A one-tailed test is appropriate here because the direction of the effect was predicted in advance on the basis of previous evidence in supervised learning (see earlier Blocking section) and in a situation where trial-specific feedback was absent (Zwickel & Wills, 2002).

Further inspection of Table 6.3 indicates that nonlearners do not show behavior consistent with blocking. This is as we would predict, because blocking should occur only if the Phase 1 categorization is learned. A contingency chi-square confirms that the proportion of blocking-consistent responses is significantly affected by performance in Phase 1, $\chi^2(1)=6.11$[1], $p < 0.05$. Taking these two analyses together, our data seem to support the conclusion that a blocking-like effect occurs in unsupervised learning. Additionally, all four participants that passed the 83% criterion in both phases made a blocking-consistent response to the test stimulus, $p = 0.06$, one-tailed, on a binomial test.

One further aspect of these results is that nonlearners appear to show nonrandom responding in the opposite direction to that pre-

TABLE 6.3
Results of the Experiment

	Learners	*Nonlearners*
Consistent with blocking	8	1
Inconsistent with blocking	2	5

[1]No corrections have been applied for the low expected frequencies of some of the cells. It has been found that even small expected frequencies do not increase the chance of type I errors (Overall, 1980). A general discussion of this issue can be found in Howell (2002, pp. 151–152).

dicted by the blocking hypothesis. Although this effect falls short of significance, $p = 0.22$, two-tailed, on a binomial test, it is interesting to speculate what may be behind this pattern. One possibility is that participants who did not learn the category structure in Phase 1 developed something of a mix of the Phase 1 and Phase 2 prototypes. These people might therefore represent the stimuli of Category 1 as derived from a prototype ABE and stimuli of Category 2 as derived from a prototype CDGF. If one accepts Pearce's (1987) assumptions about the relationship between shared elements and similarity, then test stimulus EF is more similar to ABE than to CDGF. This is because Pearce assumes that similarity is affected by the *proportion* of shared elements. EF contains one third of ABE's elements but only one quarter of CDGF's elements, so EF is predicted to be more similar to ABE than to CDGF. As a result, participants would be predicted to place EF into Category 1, which is opposite to the effect predicted by blocking.

MODELING

The results of our experiment indicate that a blocking-like phenomenon occurs in unsupervised learning. It has been our contention throughout this chapter that such an effect is predicted by our integrated model but not by Rumelhart and Zipser's (1986) competitive-learning model. In this section we show how we've supported this conclusion through computer simulations of both models.

We implemented the Rumelhart and Zipser learning rule (Equations 6.3a and 6.3b) in a network consisting of 60 input units and 2 output units. Each of the elements composing our stimuli was assigned to an input unit and the activation of that unit was set to one if the feature was present and zero otherwise. We ran the simulation 32 times, once for each participant in our experiment. For each simulation, each stimulus presented to a given participant was presented to the input units of the corresponding simulation. Stimuli were presented sequentially, and in the same order as they had been to the corresponding participant. The winning output unit for the presented stimulus was simply defined as the most active output unit. The learning rule was then applied and the next stimulus presented. The network's response to the test stimulus was coded as blocking-consistent or blocking-inconsistent using the same procedure we had used for the participants' responses (see the section Results and Discussion). Simulated participants were also categorized as "learners" or "nonlearners" using the same procedures we had used for the participants. The data from the "learners" in this simulated experiment were then assessed for the significance of the blocking effect, tested against a null hypothesis of random responding.

Equation 6.3b includes the learning rate parameter G. We therefore performed 50 simulated experiments across which G varied from 0.001 to 0.491 in steps of 0.01. None of our 50 simulated experiments produced a significant blocking effect, reinforcing our conclusion that the Rumelhart and Zipser system does not predict blocking in this experiment.

Next we replaced the Rumelhart and Zipser learning rule with our modified delta rule (Equation 6.5) and repeated the 50 simulated experiments. This time, out of the 50 runs, 42 runs were significant. The nonsignificant runs all occurred between the learning rates of 0.001 and 0.081, indicating that this prediction of our model is robust across a wide range of learning rates.

OTHER MODELS

Throughout this chapter, we've deliberately concentrated on two well-known and comparatively simple associative models of categorization that can be straightforwardly applied to the experimental procedures and stimuli we employed. In so doing, it was not our intention to suggest that the Rumelhart and Zipser or the Rescorla–Wagner models are the only, or even the best, models of unsupervised and supervised categorization respectively. In what follows we discuss the validity of some of the assumptions underlying the models we have used and consider some alternative approaches to modeling supervised and unsupervised categorization.

Stimulus Representation

The Rescorla–Wagner and Rumelhart–Zipser models both assume an *elemental* stimulus representation. For example, in our particular applications of these models we have assumed the presence of an input unit for each of the picture elements that comprise the stimuli. This kind of elemental stimulus representation can be contrasted with *exemplar* stimulus representation. In exemplar stimulus representation, each presented stimulus has its own unique representation. Exemplar models are being increasingly employed in the study of categorization and associative learning because of their proven success in very precisely modeling categorization behavior in certain circumstances (see, e.g., Nosofsky, 1986). There is also some evidence in both humans (e.g., Shanks, Darby, & Charles, 1998) and other animals (e.g., Pearce & Redhead, 1993) that appears to favor exemplar theories over comparable elemental theories. On the other hand, there are phenomena that seem difficult to explain if one assumes a purely exemplar representation but that can be easily explained if elemental representation is assumed (e.g., Gluck, 1991; Kruschke, 1996). There is also some evidence that suggests people can flexibly apply exemplar or elemental stimulus representations in response to differing task demands (e.g., Williams, Sagness, & McPhee, 1994).

It is also likely that, as suggested in General Recognition Theory (Ashby & Townsend, 1986) and stimulus-sampling theory (Estes, 1950), even two physically identical stimuli will have differing input representations due to variations in our perceptual system. The picture is yet further complicated by evidence that suggests the act of categorization can itself affect our stimulus representations (e.g., Schyns & Rodet, 1997).

Overall, it seems likely that the stimulus representations we have employed in this chapter are a simplification of the true nature of stimulus representation.

Attentional Processes

One way of thinking about the phenomenon of blocking is as a demonstration that learning is driven by surprise. In the Rescorla–Wagner theory, surprise can be thought of as directly affecting learning. This is because the learning rule states that the change in connection strength is proportional to the difference between the predicted status of the outcome (Σaw) and its actual status (λ). Some other associative theories (e.g., Mackintosh, 1975; Pearce & Hall, 1980) suggest that surprise acts indirectly through some kind of attentional process. The phenomenon of blocking does not, in itself, distinguish between these two classes of explanation. Therefore, although our discovery of a blocking-like effect in free classification indicates that unsupervised learning is surprise-driven, it does not uniquely support the Rescorla–Wagner formulation we have employed in our model. An alternative (or additional) approach would have been to add an attentional process to the Rumelhart and Zipser model.

Plasticity–Stability Dilemma

The plasticity–stability dilemma is that a system must be able to learn in order to adapt to a changing environment (i.e., it must be "plastic") but that constant change can lead to an unstable system that can learn new information only by forgetting everything it has so far learned. The back-propagation algorithm is well known to suffer from stability problems (see, e.g., the discussion of "catastrophic forgetting" in McCloskey & Cohen, 1989). Stability is also a problem for both the Rescorla–Wagner and the Rumelhart and Zipser systems.

Numerous suggestions have been made for solutions to the plasticity–stability dilemma and there is insufficient space to deal with them all here. One solution (the APECS system) is covered in chapter 7. An alternative model that is more directly applicable to unsupervised learning is Grossberg's Adaptive Resonance Theory (see, e.g., Grossberg, 1987). Adaptive Resonance Theory is somewhat related to the Rumelhart and Zipser system, but it adds a top-down process. Like Rumelhart and Zipser, input unit activity leads to one category unit being more active than the others. Unlike Rumelhart and Zipser, this category unit does not necessarily "win." It will do so only if downward connections from the category unit to the input units reproduce the input sufficiently well. If the most active unit does not predict the input very well, the same test is performed for the next most active category unit. If none of the current category units pass the test, a new category unit is created and designated the winner. In this way, the system remains able to learn about new situations while protecting what has already been learned by creating new representations to cater for the new information.

Decision Processes

In the Rumelhart and Zipser model, and in our model, the category of a presented stimulus is decided by finding the most active output unit. The process by which this happens is not specified, but is assumed to be errorless. In other words, the most active unit will always be the one that is selected. In reality, any process is likely to be imperfect and so sometimes some other unit will be selected, particularly if there are two or more units whose activations are similar. One very common way of representing this decision process is through the ratio rule. Next we consider one particular type of ratio rule, the *exponential ratio rule*.

The ratio rule compares the activity of each output unit to the activity of all other output units. In this way, a highly active output unit is selected with a higher probability if all other output units are quite low in activation than if the other units have a high activation too. In the exponential ratio rule, an additional parameter k adjusts how much influence the relative strengths of the category units have. If k is large, the most activated category is nearly always chosen. If k is small, the relative strengths of activation have very little influence on the category decision.

Formally, the exponential ratio rule is

$$Prob(category\ x) = \frac{e^{ko_x}}{\sum_{j=1}^{v} e^{ko_j}} \qquad (6.6)$$

where the activation of output unit x is represented as o_x and v is the number of output units. *Prob(category x)* is the probability that category x is selected.

The ratio rule is widely used in the modeling of categorical decision processes (e.g., Gluck & Bower, 1988; Kruschke, 1993; Nosofsky, 1986). However, there is mounting evidence that the exponential ratio rule is incorrect and that an alternative process based on the mutual inhibition of output units may be more appropriate (see, e.g., Wills et al., 2000) .

Level of Analysis

In this chapter, we've concentrated on theories that attempt to elucidate the specific processes that underlie supervised and unsupervised category learning. The two models we've looked at in most detail attempt to do this with processes that are neurally plausible. However, this is not the only approach one can take to understanding categorization. One can, for example, employ a more abstract level of analysis and consider the general problems that any categorization system must solve. In this section, we describe one theory that takes this approach—Anderson's rational model.

The core idea of Anderson's rational model (e.g., Anderson, 1991) is that, as the brain is adapted to its environment, much insight into human information processing can be gained by reflecting on the nature of information in that environment. If Anderson's model were a *normative* model, it would employ all relevant probability information about the category structure in the environment. The model is described as rational rather than normative because it also takes into account some considerations about the computational complexities of processing this information.

The model is expressed through Bayesian mathematics. For example, the probability that a person is 20 years old, given that they are a student, is quite high. This is called a conditional probability, and is expressed as P(twenty | student). Imagine you are walking down the street of a particular town and meet a 20-year-old. How likely is it that this person is a student? Another way of asking the same question is to ask for an estimate of P(student | twenty). Your estimate will, of course, be affected by P(twenty | student), but also by the overall probability that anyone is a student, P(student).

This is basically the way the rational model calculates the probabilities with which a stimulus comes from a specific category. If a new stimulus has to be categorized, the model determines the probability with which each stimulus feature would occur, given that the stimulus comes from a particular category. The probabilities for all the features are multiplied together, giving a single number for each category. The stimulus is considered to belong to the category that produces the largest number. When no prior knowledge is available, or the calculated number is below a certain threshold, a new category is created.

SUMMARY

Categorization—dividing the world into groups of things—has been studied through two basic types of experiments. The first type consists of studies where each stimulus is accompanied or followed by accurate information about category membership. These studies are by far the most common but the level of feedback given seems unlikely to be commonly available outside the laboratory. The second type of study—free classification—goes to the opposite extreme and provides no feedback whatsoever.

These two basic types of categorization experiments are reflected in the two basic types of associative categorization models: supervised models and unsupervised models. In this chapter, we concentrated on one comparatively simple associative model of each type. The first was the Rumelhart and Zipser competitive-learning system, which is a model of unsupervised learning. The second was the Rescorla-Wagner theory (aka. simple delta-rule network), which has been widely applied to human and animal data.

We went on to outline the case for integrating models of supervised and unsupervised learning. Our main argument was that only an integrated system could make use of feedback when it was available but not be paralyzed when feedback was unexpectedly absent. We then discussed one way in which competitive learning and the delta rule could be combined to create a comparatively simple integrated model.

The integrated model we proposed predicts that blocking-like effects should also occur in the absence of feedback, whereas the Rumelhart and Zipser model predicts that they will not. The results of our free-classification study suggest that blocking-like effects do indeed occur in the absence of feedback. These results therefore provide another reason for favoring our integrated model over the component models from which it was constructed.

There are a number of respects in which the model we have proposed is likely to be a simplification of a fully adequate model of categorization. These respects include our assumptions about stimulus representation, the absence of an attentional process, the absence of a process that addresses the plasticity-stability dilemma, and the absence of a realistic decision mechanism.

REFERENCES

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

Anderson, J. R. (1991). A rational analysis of categorization. *Psychological Review, 98*, 409–429.

Ashby, F. G., & Townsend, J. T. (1986). Varieties of perceptual independence. *Psychological Review, 93*(2), 154–179.

Bersted, C. T., Brown, B. R., & Evans, S. H. (1969). Free sorting with stimuli in a multidimensional attribute space. *Perception & Psychophysics, 6B*, 409–413.

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

Dickinson, A. Shanks, D. R., & Evenden, J. (1984). Judgement of act-outcome contingency: The role of selective attention. *Quarterly Journal of Experimental Psychology, 36A*, 29–50.

Estes, W. K. (1950). Toward a statistical theory of learning. *Psychological Review, 57*, 94–107.

Forster, K. I., & Forster, J. C. (2003). DMDX version 3. http://www.u.arizona.edu/~jforster/dmdx.htm

Gluck, M. A. (1991). Stimulus generalization and representation in adaptive network models of category learning. *Psychological Science, 2*, 50–55.

Gluck, M. A., & Bower, G. H. (1988). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General, 117*(3), 227–247.

Grossberg, S. (1987). Competitive learning: From interactive activation to adaptive resonance. *Cognitive Science, 11*, 23–63.

Hebb, D. O. (1949). *The organization of behavior*. New York: Wiley.

Homa, D., & Cultice, J. C. (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*(1), 83–94.

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks, 2*, 359–368.

Howell, D. C. (2002). *Statistical methods for psychology*. Belmont, CA: Duxbury Press.

Jones, F. W., Wills, A. J., & McLaren, I. P. L. (1998). Perceptual categorization: Connectionist modelling and decision rules. *The Quarterly Journal of Experimental Psychology, 51B*(3), 33–58.

Kamin, L. J. (1969). "Attention-like" processes in classical conditioning. In M. R. Jones (Ed.), *Miami symposium on the prediction of behavior: aversive stimulation* (pp. 9–31). Miami, FL: University of Miami Press.

Kruschke, J. K. (1993). Human category learning: Implications for backpropagation models. *Connection Science, 5*(1), 3–36.

Kruschke, J. K. (1996). Base rates in category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*(1), 3–26.

Mackintosh, N. J. (1975). A theory of attention: Variations in the associability of stimuli with reinforcement. *Psychological Review, 82*, 276–298.

Martin, I., & Levy, A. B. (1991). Blocking observed in human eyelid conditioning. *Quarterly Journal of Experimental Psychology, 43B*, 233–255.

McClelland, J. L., & Rumelhart, D. E. (1985). Distributed memory and the representation of general and specific information. *Journal of Experimental Psychology: General, 114*(2), 159–188.

McCloskey, M., & Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. In G. H. Bower (Ed.), *The psychology of learning and motivation* (pp. 109–165). San Diego: Academic Press.

McLaren, I. P. L. (1989). The computational units as an assembly of neurons: An implementation of an error correcting learning algorithm. In R. Durbin, C. Miall, & G. Mitchison (Eds.), *The computing neuron* (pp. 160–178). Amsterdam: Addison-Wesley.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review, 85*(3), 207–238.

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. A. (1969). *Perceptrons: An introduction to computational geometry*. Cambridge, MA: MIT Press.

Nosofsky, R. M. (1986). Attention, similarity and the identification-categorisation relationship. *Journal of Experimental Psychology: General, 115*(1), 39–57.

Overall, J. E. (1980). Power of chi-square tests for 2 × 2 contingency tables with small expected frequencies. *Psychological Bulletin, 87*, 132–135.

Page, M. (2000). Connectionist modelling in psychology: A localist manifesto. *Behavioral and Brain Sciences, 23*(4), 443–512.

Pearce, J. M. (1987). A model of stimulus generalization for Pavlovian conditioning. *Psychological Review, 94*, 61–73.

Pearce, J. M., & Hall, G. (1980). A model for Pavlovian learning: Variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review, 87*, 532–552.

Pearce, J. M., & Redhead, E. S. (1993). The influence of an irrelevant stimulus on two discriminations. *Journal of Experimental Psychology: Animal Behavior Processes, 19*, 180–190.

Rescorla, R. A. (1973). Evidence for a "unique stimulus" account of configural conditioning. *Journal of Comparative and Physiological Psychology, 85*(2), 331–338.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current research* (pp. 64–99). New York: Appleton–Century–Crofts.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 1, pp. 318–362). Cambridge, MA: MIT Press.

Rumelhart, D. E., & Zipser, D. (1986). Feature discovery by competitive learning. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 1, pp. 151–193). Cambridge, MA: MIT Press.

Schyns, P. G., & Rodet, L. (1997). Categorization creates functional features. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*(3), 681–696.

Shanks, D. R., Darby, R. J., & Charles, D. (1998). Resistance to interference in human associative learning: Evidence of configural processing. *Journal of Experimental Psychology: Animal Behavior Processes, 24*(2), 136–150.

Stone, G. O. (1986). An analysis of the delta rule and the learning of statistical associations. In D. E. Rumelhart, J. L. McClelland, & the PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 1, pp. 444–459). Cambridge, MA: MIT Press.

Werbos, P. J. (1974). *Beyond regression: New tools for prediction and analysis in the behavioral sciences.* Unpublished doctoral dissertation, Harvard University, Cambridge, MA.

Widrow, B., & Hoff, M. E. (1960). Adaptive switching circuits. *Institute of Radio Engineers, Western Electronic Show and Convention, Convention Record*, (Part 4, pp. 96–104).

Williams, D. A., Sagness, K. E., & McPhee, J. E. (1994). Configural and elemental strategies in predictive learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*, 694–709.

Wills, A. J., & McLaren, I. P. L. (1998). Perceptual learning and free classification. *The Quarterly Journal of Experimental Psychology, 51B*(3), 235–270.

Wills, A. J., Reimers, S., Stewart, N., Suret, M., & McLaren, I. P. L. (2000). Tests of the ratio rule in categorization. *The Quarterly Journal of Experimental Psychology, 53A*(4), 983–1011.

Zwickel, J., & Wills, A. J. (2002). Is competitive learning an adequate account of free classification? In W. Gray & C. Schunn (Eds.), *Proceedings of the 24th Annual Conference of the Cognitive Science Society* (pp. 982–987). Mahwah, NJ: Lawrence Erlbaum Associates.