Running head: HEBBIAN LEARNING

How Far Can You Go with Hebbian Learning, and When Does it Lead you Astray?

James L. McClelland

Center for the Neural Basis of Cognition and Department of Psychology, Carnegie Mellon University,

Pittsburgh, PA.

115 Mellon Institute, 4400 Fifth Avenue, Pittsburgh, PA 15213

Phone: +1-412-268-3157; Fax +1-412-268-5060; email jlm@cnbc.cmu.edu

Abstract

This paper considers the use of Hebbian learning rules to model aspects of development and learning, including the emergence of structure in the visual system in early life. There is considerable physiological evidence that a Hebb-like learning rule applies to the strengthening of synaptic efficacy seen in neurophysiological investigations of synaptic plasticity, and similar learning rules are often used to show how various properties of visual neurons and their organization into ocular dominance stripes and orientation columns could arise without being otherwise pre-programmed. Some of the pluses and minuses of Hebbian learning are considered. Hebbian learning can strengthen the neural response that is elicited by an input; this can be useful if the response made is appropriate to the situation, but it can also be counterproductive if a different response would be more appropriate. Examples in which this outcome-independent Hebbian type of strengthening might account at least in part for cases in which humans fail to learn are considered, and computational models embodying the Hebbian approach are described that can account for the findings. At a systems level, Hebbian learning cannot be the whole story. From a computational point of view, Hebbian learning can certainly lead one in the wrong direction, and some form of control over this is necessary. Also, experimental findings clearly show that human learning can be affected by accuracy or outcome feedback. Several ways in which sensitivity to feedback might be incorporated to guide learning within a fundamentally Hebbian framework for learning are considered.

How Far Can You Go with Hebbian Learning, and When Does it Lead you Astray?

Connectionist models that learn using error-correcting learning rules have long played a role in the exploration of issues in cognitive and linguistic development. In my own work in the late 1980’s and early 1990’s, I explored the idea that predictive error-driven learning provided the engine that drives cognitive development (McClelland, 1994). The central idea was that the developing child is continually making implicit predictions about the future state of the world based on current inputs, using an internal model embodied in a connectionist network. Any mismatch between the child’s expectations and observed events provides an error signal, indicating that the child’s internal model must be updated. Learning then occurs by adjusting the parameters of the mind — the connection weights in the network — to reduce the discrepancy between predicted and observed events. The necessary changes are determined by the use of the back-propagation algorithm (Rumelhart, Hinton, and Williams, 1986).

I applied this predictive error-driven learning approach in a series of models (McClelland, 1989; McClelland & Jenkins, 1991; McClelland, 1995) addressing aspects of cognitive development as revealed by the work of Siegler (1976) on the Inhelder and Piaget (1958) balance scale task. Other models that use back propagation to adjust connection weights to reduce the difference between predicted and observed events include Elman’s (1990) model of the acquisition of syntax through learning to predict the next word from previous words, and St. John and McClelland’s (1990) model of learning to comprehend sentences through learning to predict the characteristics of the events described by sentences. The approach has also been applied to the development of conceptual and physical knowledge during infancy (Munakata et al, 1997) and to many aspects of conceptual development during childhood (Rogers and McClelland, 2004). Several papers in this volume present ongoing investigations that can be thought of as exemplifying this approach.

In spite of my own continuing reliance on error-driven learning (especially in the work with Rogers, cited above) I have recently begun exploring learning that can occur in the absence of an error signal (McClelland, 2001). Specifically, I have been exploring the possibility that Hebb’s famous proposal for learning in neural systems may provide some guidance in addressing aspects of human learning – particularly some of its failures as well as its successes – that may not be fully addressed within the error-driven approach. In the course of this work I have also had occasion to begin to think more generally about the fact that we learn from our own reactions and behaviors, as well as from the sequences of events that we see in the world. The key questions addressed in this chapter are:

> How well do networks based on Hebbian learning work as computational systems, and when do they fail?
>
> How do the successes and failures of Hebbian systems compare to the successes and failures of human learning?

In the first part of this chapter, I will review Hebb's proposal for learning and describe briefly some of the neuroscience evidence that has supported the basic proposal and led to a family of Hebb-like learning rules used widely in biologically-oriented models of neural network learning. The second part of the chapter will explore some puzzling findings concerning successes and failures of learning that can be addressed with models in which learning occurs on the basis of these biologically-oriented learning rules. In the course of this we will encounter some evidence that makes it clear why the Hebbian approach is not fully sufficient without some elaboration to encompass sensitivity to outcome information. This will lead to a brief discussion of how such information may be accommodated within a fundamentally Hebbian framework. I will conclude by returning to the broader question of how we learn from our own behavior. I will suggest that even here Hebbian learning has its limits, and consider ways it may be regulated by internally-generated signals that can be used to help guide our learning in the right directions.

Hebb's proposal and the biology of Hebbian learning

In a famous passage in his 1949 book, Hebb proposed a neural mechanism for learning:

> 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.

This passage captures the fundamental idea rather well, but is somewhat cumbersome to process. A simpler, catchier wording captures the key idea quite succinctly:

> Cells that fire together wire together.

The source of this latter version is unknown, but it is oft-repeated in discussions of Hebbian learning in neuroscience.

Hebb's proposal has been the inspiration of a wide range of biological research on long-term potentiation, usually conducted in slices taken from the hippocampus of the rat. Long-term potentiation is explored by providing a train of impulses to a bundle of fibers projecting to a target neuron (or group of neurons) whose state of activation is being monitored (and potentially directly controlled). When the train of impulses is strong, so that the post-synaptic neurons fire, the efficacy of smaller test pulses arriving on the bundle of fibers is increased, or potentiated. Typically there is a short-lasting component that decays over minutes, to leave a longer-lasting increment that can persist for the lifetime of the preparation; it is this longer-lasting increment that we refer to with the phrase long-term potentiation (LTP). It is well-established that a weak input, itself insufficient to produce LTP, could result in potentiation if paired with a stronger input arriving on other fibers (Barrionuevo and Brown, 1983; McNaughton, Douglas, and Goddard, 1978).

Research on LTP and related phenomena is ongoing, and several important additional facts have been established. One of these, the phenomenon of spike-time-dependent-plasticity (STDP), further supports Hebb's proposal. What STDP refers to is the finding that the exact timing of neural firing in pre- and post-synaptic neurons can influence LTP. Indeed, under some conditions, LTP is maximal if the sending neuron fires just before before the receiving neuron fires, which it must if it is to participate in firing the receiving neuron; and LTP is reversed if sending neuron fires after the receiving neuron fires. This phenomenon of STDP arises when individual pre-and-post synaptic events are considered at a fine time grain. The role of spike-time-dependence in the living, behaving organism is debated, since the timing of spikes tends to be quite variable; but STDP can help neurons become more selective for temporal coincidences (Abbott & Nelson, 2000; Song, Miller, & Abbott, 2000), and is likely to receive considerable further exploration as a mechanism for learning fine-grained temporal information.

In any case, most existing models aimed at system-level and behavioral phenomena, including all those to be considered here, do not address neuronal activity at this fine level of granularity. Instead such models use continuous activation values, thought to reference the mean firing rate of a population of neurons (generally it is assumed that a population is required to provide a sufficient signal to have an impact on receiving neurons). At this level, LTP is generally found to be well-captured as depending on the product of separate functions of pre- and post-synaptic activity:

$$\Delta w_{rs} = \varepsilon f_r(a_r) f_s(a_s) \tag{1}$$

Learning rules obeying relations of this form will be considered Hebbian; in the simple case where the functions are identify functions ($f_r(a_r) = a_r,\ f_s(a_s) = a_s$) and where a_r and a_s take on activation values of 1 or 0, this learning rule captures the simple “cells that wire together fire together” idea in a straightforward and simple form. However, in this form the rule is not quite adequate, either to capture findings from LTP studies or to address issues in development and learning.

Two findings from LTP research have contributed to the specification of the functions f_r and f_s. (1) Early research by Levy and Steward (1979) established that LTP is accompanied by a corresponding depression of the efficacy of inputs that are not activated when the receiving neuron is activated. This phenomenon, known as “hetero-synaptic long-term depression” can be captured in models by assuming that the pre-synaptic function f_s in the equation above is given by the difference between the sending unit’s activation and the existing value of the connection weight:

$$f_s(a_s) = (a_s - w_{rs}) \tag{2}$$

(2) Under conditions where strong trains of input stimulation produce LTP, repeated weak trains tend to produce a long-term depression or reduction of synaptic efficacy (Dudek & Bear, 1993). This phenomenon was predicted from a model of developmental stabilization in which the Hebbian update equation is assumed to be non-monotonic (Bienenstock, Cooper, & Munro, 1982). With very low levels of activity in the receiving neuron, a_r, the value of the post-synaptic term is 0; as activity increases it first becomes negative, then as activity increases further still it becomes positive. This non-monotonic function, called ϕ, was proposed because it tends to encourage the stabilization of neuronal responses and to promote sharp

tuning of the receiving neuron's receptive field. This occurs because inputs that only weakly activate the neuron are weakened still further. Although the use of the ϕ function would likely improve most models, the slightly simpler form below, in which $\phi(a_r)$ is approximated simply by a_r, is used in many computational models (Grossberg, 1976; Rumelhart & Zipser, 1985).

$$\Delta w_{rs} = \varepsilon a_r (a_s - w_{rs}) \tag{3}$$

A variety of different names have been given to this learning rule. Although the name is a bit cumbersome, I will refer to this rule as the Hebbian redistribution rule or the Hebb$_r$ rule for short, since it re-distributes the incoming weights to a unit so that they are aligned with the input pattern activating the unit.

Applications of Hebbian Learning in Biological and Psychological Development

Insert Table 1 about here

Learning rules of the general form of Equation 1 have been used in many models addressing aspects of biological and psychological development. On the biological side, Linsker (1986a, 1986b, 1986c) has used a variant of Equation 1 to model the emergence of center-surround receptive fields, oriented edge detectors, and ocular dominance columns during pre-natal development. Considering first center-surround receptive fields, what Linsker observed is that the post-synaptic or receiving neuron tends to strengthen the subset of its inputs that are most highly correlated with each other. To see why this should be, suppose there are three inputs to a receiving neuron, and suppose that the first two occur together but the third occurs when the other two do not (this very simple case represents the extreme of positive vs. negative correlation; the same principle applies to less extreme cases). As illustrated in Table 1, the post-synaptic term in the Hebbian learning rule will be larger, on average, when either the first or the second input is active than it is when the third is active, so Equation 1 will cause the weights to the receiver from the first and second units to increase more than the weight to the receiver from the third unit. This tendency, when opposed by an overall tendency for connection weights to decay, can result in a situation where the inputs that have relatively strong correlations with other inputs to the receiving unit end up with connection weights to the receiver that grow stronger over time, and those that are less correlated with the other inputs to the receiver end up with weights that grow increasingly more negative.

Linsker (1986a,b) considered a network consisting of several two-dimensional arrays of cells (similar to the retina and subsequent layers in the visual pathway), with the density of connections reaching a given cell in one layer from the preceding layer obeying a decreasing (Gaussian) function of distance between the positions of the cells in the two layers. He then considered what would happen if the cells in the first layer fired spontaneously and independently. Because of the distance-dependence in the initial connections, neighboring cells in the second layer are likely to receive connections from some of the same cells in the first layer, so that the inputs — and therefore the resulting activations — of these neighboring cells will be slightly correlated; the degree of correlation falls off with distance between the neighbors. Now, the crucial observation concerns the inputs to cells at the third layer. Each third-layer cell receives relatively dense projections from the cells at the corresponding location in the preceding layer, with density falling off with distance. As a result, cells near the center of the projection to the third-layer cell tend to have a greater correlation with all the other cells projecting to the third layer cell than those near the edges. The result is that the connections from cells near the center of the projection end up with positive connections and those near the edges end up with negative values, thereby producing center-surround receptive fields. If the outputs of center-surround cells at the third layer are propagated forward through several further layers, Linsker (1886b) found that edge detectors can emerge. The reasons for this are less intuitive but still follow from the principle that the neurons at higher levels are maximizing their responsiveness to maximally correlated clusters of cells in their inputs.

In order to organize the edge detectors that emerge at higher levels into orientation columns, Linsker (1986c) assumed a particular “cortical interaction function” in which nearby neurons tend to excite each other and neurons at a greater distance tend to inhibit each other. The result of this is that neighboring neurons tend to encourage each other to have similar response properties, while neurons with slightly greater separations tend to discourage this tendency. Orientation columns thus arise as the bi-product of Hebbian learning coupled with the local excitatory and mid-range inhibitory interactions. Similar principles (most importantly, the assumption the neighboring neurons in the same eye are more correlated during spontaneous pre-natal activity than neurons in two different eyes) result in the emergence of eye-selective neurons and ocular dominance columns where inputs from the two eyes converge (Miller et al, 1989).

It should be noted that there is now evidence that some aspects of cortical neural organization are laid down during early phases of development, before the developing neurons become active (Crowley & Katz, 2000). This organization may depend on chemical gradients and other mechanisms that channel axonal growth, providing an initial coarse framework that encourages map-like spatial organization and alternating stripes of innervation from the two eyes. Thus, by no means is it assumed that all aspects of neural organization depend on Hebbian mechanisms. What is widely assumed is that a very basic framework provided prior to the onset of activation is further refined and maintained by Hebbian learning mechanisms.

Insert Figure 1 about here

A particular network architecture that has proven fairly popular in models both of psychological and biological development is shown in Figure 1. The particular network shown, based loosely based on the proposals of Kohonen (1982), provides two two-dimensional sheets of neurons, corresponding perhaps to a sensory surface (on the retina or the skin) called the input layer and a second layer of units called the representation layer that receives projections from the input surface. As above, initial weak connections are assumed to have a weak topographic bias, so that a neuron at a given location in the input layer will tend to have slightly stronger connections to corresponding locations in the representation layer.

Within each layer, short-range excitatory connections and longer-range inhibitory connections are assumed to enforce a scenario in which a localized blob of activity at the input layer gives rise to a corresponding localized blob of activity at the receiving layer. (In Kohonen’s idealization of this, the receiving unit with the strongest weighted input is simply chosen as the winning unit. It and its neighbors are assigned activation values ranging from 1 to 0, following a Gaussian falloff with distance from the winning unit. Similar architectures have frequently been used to model refinement of map-like cortical representations (e.g., Goodhill, 1993).

In addition to modeling structural features of the underlying neural substrate, models similar to the one in Figure 1 have also been used extensively in models of development at a behavioral level. Two cases in point are the models of category formation introduced by Schyns (1991) and the model of infant's sensitivity to causal even sequences introduced by Cohen, Chaput, and Cashon (2002). In both models, the tendency of such map-like representations to group similar inputs together and to learn to respond to patterns of coherent covariation in these inputs plays an important role in the accounts offered for important aspects of cognitive development.

Computational Limitations of Hebbian Learning

Hebbian learning is relatively weaker than error-correcting approaches, where there is an explicit measure of the network's error and an explicit procedure for adjusting weights so as to minimize the error. Such error correcting learning algorithms have been very successful in allowing networks to solve hard problems, and in so doing to discover useful representations that aid in their solution. Networks trained using strictly Hebbian approaches are not really constrained in this way. Thus, the computational question, "How well do networks based on Hebbian learning work as computational systems?" is answered by many modelers "not very well at all". However, it may be worth noting that it is possible to guide some types of networks that are trained with the Hebb$_r$ rule in Equation 3 to some extent. To see this, let us consider what happens in a simple competitive network of the kind used by Rumelhart and Zipser (1985). Competitive networks consist of two layers of units, as in the Kohonen model, but there are no lateral spread of activation: When an input is presented, one of the representation units is chosen as the winner and is allowed to adjust its connection weights according to Equation 3; the weights coming in to all other units are left unchanged. In such a network, each competitive unit tends to pick out its own cluster of similar inputs, a simple form of category formation. Each unit's weight vector ends up positioned at the centroid of its cluster, and the assignment of patterns to clusters tends to maximize within-cluster similarity. Thus the approach provides a simple, input driven category-formation algorithm.

However, sometimes the similarity structure in a set of inputs isn't a sufficient guide to their correct categorization, as defined by environmental contingencies; at the very least, some aspects of similarity may be more important than others, for example. Figuring out how to weight different kinds of similarity is one of the strengths of error-correcting learning, and indeed, in such models, the similarity structure over inputs can sometimes be completely remapped at hidden layers using error correcting learning. The point I would like to emphasize here, however, is that Hebbian learning mechanisms are not completely insensitive to guidance by such things as category labeling information (or any information reliably correlated with variable item properties). Specifically, within the Rumelhart and Zipser model, the clustering process can be influenced by providing additional input, such as a category label, along with each pattern. The label input is treated as just another part of the input pattern, and as such it tends to make the patterns that are assigned the same label more similar to each other, and those assigned different labels less similar, thereby altering the clustering of inputs. This method was used extensively to cause networks to group divergent inputs together in the simulations of Rumelhart and Zipser (1985) and a similar approach was used by Schyns in his Kohonen-network based model of perceptual category learning with and without labeling information.

My point here is not to suggest that Hebbian learning is sufficient to solve all hard learning problems. I would suggest, however, that the potential of this approach may be underestimated. Later in the paper I will turn to other ways in which an inherently Hebbian learning process may be guided. For now, however, I consider some characteristics of Hebbian learning which may help explain some successes and failures of learning in experiments with human subjects.

Some Psychological Pluses and Minuses of Hebbian Learning

A key observation about Hebbian learning is that it tends to strengthen the pattern of neural response that occurs to a particular input. When such responses are desirable, their strengthening can lead to increased accuracy, fluency, etc., of the perception, emotion, thought or action associated with this neural response, and this may well underlie the increased fluency of performance that occurs with practice in many information processing tasks.

However, when the response that is evoked by an input is not desirable, its strengthening may have deleterious consequences. One such deleterious consequence is the phenomenon of dystonia that can occur in musicians and writers. Dystonia appears to reflect a tendency for the various digits of a person's hand to loose their independence and often to enter into a state of chronic activation experienced as a cramp. Although the details are not fully clear about how these disorders develop, they do occur in individuals who persistently activate a set of muscles together, e.g. by gripping the neck of a violin or a writing implement with several digits simultaneously. This continual co-activation of muscles (and the neurons that drive them) could lead to Hebbian strengthening of synaptic connections among the neurons that control the muscles, such that the intent to move any one muscle would then lead to co-activation of all of them. Once set up, the dystonia may be very difficult to correct, especially in musicians who must continue to perform. A second kind of deleterious consequence could occur in phobias and in racism. If you react in fear upon seeing a snake or a person from a particular racial background, Hebbian synaptic strengthening may increase the tendency to respond in this way, even if nothing bad actually happens. A third kind of case may arise in persons—including scientists—who have long practiced a particular way of thinking, and are then confronted with an entirely new and incommensurate pattern of thought. The idea that such entrenched "habits of mind" may provide barriers to the discovery and acceptance of new scientific ideas is discussed at length in two books by Howard Margolis (1987, 1993) a philosopher and historian of science. For example, Margolis discusses nightmares Darwin had for years in which he was plagued by a vision of the complex structure of the human eye. The entrenched habit of thinking of the eye as too complex to arise without divine design was difficult apparently even for Darwin himself to shake.

In the remainder of the present article, I will provide an update on investigations of two other cases in which Hebbian learning may provide at least part of the explanation for failures of human learning, and where experimental evidence now suggests that procedures developed with the strengths and weaknesses of Hebbian learning in mind can allow us to arrange conditions so that learning will be more successful. Both puzzles arose within the context of the complementary learning systems model of McClelland, McNaughton and O'Reilly (1995). According to this model, there is a gradual or slow learning system, in the neocortex, which subserves the acquisition of language, conceptual knowledge, cognitive skills, and many other sorts of cognitive abilities that are acquired gradually in the course of development. This system is complemented by another, fast learning system, in the hippocampus and related structures in the medial temporal lobes, that allows for the rapid formation of arbitrary associative conjunctions such as the conjunction of a name with a face or the particular arbitrary combination of elements that together form an experienced episode. This theory continues to provide a guiding framework for my own thinking about learning and memory in the brain, but a few years ago two things began to puzzle me.

Paradox of New Learning In Amnesia

One of these puzzles arose in thinking about new learning in amnesia. The amnesic syndrome, as seen in patient HM (Scoville & Milner, 1957), results from removal or severe damage to the hippocampus and related structures, and leaves the cortical learning system intact, which, on the theory, is capable of gradual learning. And indeed, there is evidence of some slow learning in patients like HM. For example, Milner et al (1968) had shown that HM was able to acquire some new declarative information. He was able to identify John F. Kennedy from his profile on a coin, and he could report that Kennedy had been the president and that he had been assassinated. Kennedy's election and assassination both occurred after HM became amnesic, so HM must have learned this information. Furthermore Gabrieli, Cohen, and Corkin (1988) report evidence that HM learned new words that entered the language after he became amnesic. Though his acquisition was certainly not normal, it was nevertheless clear that he had acquired the meanings of some new words. However, in a laboratory experiment, Gabrieli et al. made absolutely no progress in teaching HM the meanings of eight infrequent words. Several approaches were used to try to get HM to learn the words. The main one involved showing a definition and giving HM a chance to choose which of the eight words went with it. A response was required on every trial, and HM had to keep trying until he chose the correct answer. The answer was then eliminated as a possibility, and the definition of another word was then given. HM then had to choose from seven alternatives. This process iterated until a single choice remained. HM practiced this task extensively over a several week period, but never made any progress. At the end of the experiment he could define correctly only one of the words, a word he happened to know before the beginning of the experiment. Other studies with HM reviewed in Milner et al (1968) indicated that he was unable to learn even very short lists of arbitrary paired associates. I was puzzled by these failures, because I expected HM to be able to learn new things, albeit gradually, via his largely intact neocortical learning system, and because of the evidence that he had indeed learned some new things during the time he was amnesic.

While one could envision a host of reasons for the apparent discrepancy indicated here, one possibility that occurred to me was as follows. Perhaps the discrepancy reflects a special disadvantage that would accrue in learning new material if a patient who is only able to learn slowly were forced to make a response on every learning trial. According to the complementary learning systems model, residual cortical learning is expected to be very gradual, depending on the cumulation of very slight connection adjustments on each learning opportunity. But many experiments (including paired-associate learning tasks and the version of the word-learning task used by Gabrieli et al) combine a test of memory with every learning trial. Since the amnesic patient will have learned very little from the first exposure to the correct answer, the test of memory that occurs on the second exposure is very likely to result in an incorrect response. If this incorrect response is strengthened (as it would be under the Hebbian approach), this would tend to work against correct elicitation of the desired association. On subsequent trials this same process would continue. Depending on details (e.g. the likelihood of the same interfering response being repeatedly elicited by the test, etc), this could result in a complete failure to show improvement, even though connection strengths are being adjusted on each learning trial. Possibly, connection weights promoting both the correct and the incorrect response would be strengthened on every learning trial. While strengthening the weights promoting the correct response would tend to lead to improvement, strengthening the weights promoting an error would tend to increase interference that would counteract the improvement.

Prediction and Experiment

An obvious prediction follows from this account: A procedure in which the patient is given repeated exposure to a set of cue-target pairs without requiring any response during exposure should be more likely to result in correct responding than a procedure that attempts to elicit a response from each cue alone, prior to exposure to the correct response. For future reference in this article, we call the former procedure a study-only procedure, while we call the second procedure a generate-study procedure. The standard procedure used in paired associate learning, often called the method of anticipation is of course an example of a generate-study procedure, and the multiple-choice procedure of Cohen et al is similar to such a procedure in that it tends to result in the elicitation of many errors before exposure to the correct word-definition association.

Once this prediction became apparent, I quickly discovered several studies that have reported advantages for amnesic patients with the study-only procedure (Baddeley and Wilson, 1994; Hamman and

Squire, 1995; Hayman et al, 1993). The Hayman et al study is perhaps the most dramatic. The investigators tested patient KC (an individual who became profoundly amnesic due to a closed-head injury arising from a motor-cycle accident) with what might be called amusing word definitions, similar to some crossword puzzle cues. KC was given repeated exposure to these materials to see if he would learn the experimenter's target response for each clue. Two examples were “a talkative featherbrain” (parakeet) or “Marlon Brando's wife” (godmother). Materials were divided into those for which KC had a pre-existing (incorrect) association and those for which he drew a blank, and half of each type were presented repeatedly using the study-only procedure, while the other half were presented using the generate-study procedure. Items were further subdivided into those studied once per session and those studied twice per session. When tested after several study sessions with all of the materials, KC showed a clear study-only advantage, and the results were most dramatic for the case in which KC had a pre-existing response. Among items presented only once per session, KC showed no progress in the generate-study condition, and he persisted in giving the same response he had generated himself in most cases. In contrast he learned to produce the experimenter's response on 67% of the cases in the corresponding study-only condition (items for which he had a pre-existing response, presented once per session). This finding is highly consistent with the idea that elicitation of incorrect responses results in their strengthening, thereby either blocking learning of the correct cue target association or perhaps simply masking evidence of learning of the experimenter-defined cue-target association by maintaining the pre-existing response at sufficient strength to compete with the experimental target item.

The findings in the Hayman et al experiment are supportive of an important role for Hebbian strengthening of incorrect responses. However, other studies have produced much smaller study-only advantages compared to a generate-study condition. For example, Hamman and Squire (1995) used cue-target materials that form meaningful (but improbable) sentence-like units, such as “Medicine cures HICCUPS”. Here “Medicine cures” is the cue and “HICCUPS” is the to-be-learned target response. Note that the response words were chosen to be unlikely but meaningful, and were pre-tested to ensure that they would be generated very infrequently to the cue prior to first exposure (<5%). While Hamman and Squire did find a study-only advantage in this experiment, the effect was not large. Furthermore, the size of the effect was correlated with the degree to which the patient exhibited signs of a frontal deficit, and thus it was suggested that the study only advantage might not be relevant to understanding failures of learning in relatively pure cases of amnesia.

In this context I was lucky to have the opportunity to re-visit the issue in a collaborative study (Bird, Cipolotti, Kwok, & McClelland, 2004) of patient VC, a severely amnesic patient that has been previously characterized by Cipolotti et al. (2001) as having severe retrograde as well as anterograde amnesia as a result of a hypoxic incident. VC’s damage appears to be restricted to the medial temporal lobes and he shows no signs of frontal deficits. As in the Hamman and Squire study, we used improbable but meaningful sentence-like materials. The study phase was not presented to VC as a memory experiment; rather he was told for the study-only (SO) items to listen to each item as it was read by the experimenter, repeat it, and then rate it for meaningfulness. For the generate-study (GS) items he was first given the cue, asked to produce his own completion (which could, for example, have been “INFECTION” for the “medicine cures ...” example) and then to listen to the experimenter’s version, repeat it, and provide a meaningfulness rating. Thus our procedure ensured that a response was elicited on every generate-study trial, something that was not done in the Hamman and Squire experiment. The 20 GS and 20 SO items were presented 8 times each in interleaved sequence (GS items alternated with SO items), and an additional 20 unstudied control items were included only at test. It may be worth noting that there is a contrast in the framing of the task for subjects between the Hamman and Squire experiment on the one hand and the present study (along with Hayman et al) on the other. In Hamman and Squire, the subjects appear to have been instructed to try to learn the experimenter’s target word, as in the standard method of anticipation used in paired-associate learning experiments. This presumably engenders an effort to try to retrieve an episodic memory of the previously presented cue-target pair. In contrast, our study and Hayman et al investigated incidental learning based on repeated encounters with the cues, self-generated responses, and target materials.

Combining the findings from two runs of the experiment produced very clear-cut results: When tested two minutes after the end of the study phase, VC produced the correct response for 45% of the study-only items and only 15% of the generate-study items. The difference was highly reliable, and the advantage for the generate-study condition compared to the control condition (5% correct responses) was not significant. This finding thus confirms that elicitation of incorrect responses leads to strengthening of the elicited responses that can prevent or mask learning in a generate-study condition, and it is clearly consistent with the basic Hebbian notion that we strengthen whatever response we make to a given input.

It should be noted that a study-only advantage is rarely found in normal subjects (but see Baddeley and Wilson, 1994). Indeed, Hamman and Squire (1995) found an advantage for generate-study in several normal control conditions, including one that equated control performance with amnesic performance using a long delay between study and test. This may reflect the fact that in their study, normal subjects approached the generate phase of each study trial (excluding the very first study trial, of course) as an opportunity to recall the correct response. With the aid of their intact hippocampal system, the controls would then have been able to reinstate the correct experimenter-defined response in many cases, thereby reducing the likelihood of incorrect response generation, and providing the correct answer to the cortex so that this response, rather than some other incorrect one, is strengthened by Hebbian learning. A computational model based on these ideas has been used to simulate the results of several of the experiments discussed in this section (Kwok, 2003).

Why Can't Japanese Adults Learn to Distinguish /r/ and /l/?

A second puzzle that I contemplated in the context of the complementary learning system model concerns the apparent loss of plasticity for some aspects of language learning in adulthood. The question here was put to me by Helen Neville (personal communication): If the slow learning system in the cortex is as generally applicable to all forms of learning as I assume it to be, and if this system remains capable of new learning in adulthood, then why is it that acquisition of the phonology and syntax of a non-native language appears to be so hard for adults to learn? In fact it should be noted at the outset that the facts remain somewhat unclear on several relevant points. While some have emphasized differences in degree of mastery of subtle syntactic and semantic aspects of a language as a function of age of arrival in the new language context (Johnson & Newport, 1989), others have emphasized that there is in fact gradual acquisition even of some of the most difficult aspects of second languages, and have pointed out that age of arrival may be negatively correlated with degree of emersion in the new language culture; on this view, one account of the difficulty of later arrivals is that they simply have less exposure. The exact extent of reduced plasticity for language acquisition remains unclear, and accounts other than the one below are possible (see Seidenberg and Zevin, this volume). Nevertheless, it occurred to me that the deleterious consequences of Hebbian learning might be a contributing factor, strengthening undesirable responses that interfere with new learning, just as with the failure of learning under generate-study conditions in amnesia.

Consider the case of Japanese adults confronted with the need to perceive the distinction between the English speech sounds /r/ and /l/. In their native language, Japanese speakers do not distinguish these sounds. The do have a sound often called “the Japanese tap” that sounds to the English ear sometimes like an /r/ or an /l/, though it also bears some similarity to the reduced form of /d/ and /t/ that occurs intravocallically in “cider”. Suppose that Japanese speakers hear both English /r/s and /l/s as examples of their native tap sound, and that this perception is, like that of other speech sounds, categorical, in that they have the very same percept for sounds they treat as belonging to the same category. Then if Hebbian learning is occurring, it may unhelpfully strengthen the tendency of /r/ and /l/ sounds to elicit this categorical percept. Note that under this account, the mechanisms of plasticity would be completely intact, but they would be unhelpfully working against the acquisition of a discrimination between /r/ and /l/.

My collaborators and I have developed two computational models illustrating how this process may work, using a Hebb-like learning rule like the one in Equation 3. The first of these (McClelland et al, 1999) uses a Kohonen network architecture like the one shown in Figure 1, while the second (Vallabha and McClelland, 2004) uses a three-layer recurrent neural network with attractor dynamics. The second model has several advantages over the first, but basically illustrates the same process in action. The second model has also been used to provide a simulation of many aspects of the data we obtained from an experiment designed to explore the implications of this Hebbian account for methods that might help us teach Japanese adults to hear the difference between the English /r/ and /l/ sounds. The studies partially supported the Hebbian account but also indicated that it is incomplete in its simplest form, so I consider the experiments before returning to further discussions of the modeling work.

Implications for Teaching /r/-/l/ Discrimination

Our experiments (McCandliss et al, 2002) revolved around the following simple idea: If the Hebbian explanation is correct, it should be possible to help Japanese adults learn to distinguish /r/ and /l/ using versions of /r/ and /l/ stimuli that exaggerate the distinctions between them to the point that they will elicit distinct percepts for Japanese listeners. Hebbian learning would then operate to strengthen the tendency of each input to elicit a distinct percept. We might then be able to gradually reduce the difference between the stimuli, as long as we continue to ensure that they generally elicit distinct percepts, so that gradually, the listeners would come to perceive natural, unexaggerated tokens of /r/ and /l/ as different. Note that, from a Hebbian perspective, no outcome information should be necessary to allow listeners to learn to hear the /r/ and /l/ stimuli differently. Thus no information was given about response accuracy in our first experiment.

Insert Figure 2 about here

To construct the stimuli for our studies, we took two minimal pairs — “road-load” and “rock-lock” — and used them to construct 60-step stimulus continua ranging from an exaggerated /r/ to an exaggerated /l/ (See Figure 2). A male native English speaker produced each word carefully several times, and examples of each pair were chosen such that the articulations could be aligned to each other over time. For each continuum, the shared word body (“_oad” or “_ock”) was the same for all members of the continuum. The onset of each item was constructed from the set of coefficients obtained by linear predictive coding of the the natural /l/ and /r/ stimuli. Both continua were pre-tested with native English speakers to ensure that exaggerated stimuli were still identifiable as /l/ and /r/ and to determine the position of the boundary between the /l/ and /r/ percepts and also to identify the edges of the gray zone between the /r/ and /l/ categories for English speakers.

For our first experiment, we contrasted an adaptive training procedure that started with initially exaggerated /l/ and /r/ stimuli with a fixed training procedure using the stimuli just at the edge of the gray zone between the /l/ and /r/ percepts for native English speakers (See Figure 2). In the adaptive condition, training began with exaggerated stimuli falling outside the native English range and spaced an equal number of steps on either side of the native category boundary. Eight Japanese adults received training in each condition. Only subjects performing below 70% correct in a pretest of their ability to correctly identify the fixed training stimuli were included in the experiment. Half of subjects in each group received training with “rock-lock”; the other half received training with “road-load”.

The training procedure was very simple. One each trial, the /r/ or the /l/ stimulus from the training continuum was presented. Subjects responded by pressing a key to choose /r/ or /l/. No accuracy feedback was given. In the adaptive condition only, stimuli were adjusted between trials based on the subject’s performance. After each error, one of the two stimuli was replaced by the next easier (more exaggerated) token, alternating which item was adjusted to maintain symmetry around the native boundary. After eight successive correct responses, one of the two stimuli was replaced by the next harder token, again alternating to maintain symmetry. Training took place over three 20-minute sessions, each involving 480 training trials and 50 probe trials (with the fixed training stimuli, so that performance of the two groups could be directly compared). Half of subjects in each group continued for three additional sessions.

Insert Figure 3 about here

The results of the experiment were very clear-cut, and consistent with the Hebbian analysis (Figure 3, top panels). All eight subjects in the adaptive condition showed an improvement in their identification of the stimuli on the trained continuum after three sessions of training (top left panel). There was only a slight improvement for the fixed training group overall (top right), and the amount of improvement was no larger than that seen in a control group (not shown in the figure) of eight additional subjects who received the same pre- and post-testing separated by the same number of days as the adaptive and fixed training groups, but had no training between tests.

The findings of our first study suggest that there is considerable residual plasticity in the phonological systems of Japanese adults. Their failure to learn under normal conditions may reflect not so much a loss of plasticity as a tendency for the mechanisms of learning to maintain strongly established perceptual tendencies, as expected under the Hebbian analysis.

Results of our experiment are also consistent with the predictions of a Hebbian account of perceptual learning: Successful learning can occur without outcome information. If stimuli are exaggerated so that distinct percepts are produced, learning occurs. Learning is far less successful using the fixed training stimuli. However, it should be noted that several of the subjects in the fixed training condition did eventually begin to learn the /r/–/l/ distinction. It should also be noted that there were very large individual differences in the learning progress of subjects in the adaptive training condition. One thought about this, based on an analysis of learning in the recent model of Vallabha and McClelland (2004), revolves around the idea that the subjects should be viewed as falling on a continuum in terms of their initial tendency to hear the /r/ and /l/ stimuli as the same or different. This tendency occurs in the model because the perceptual representations formed in the model are patterns of activity that exhibit attractor-like properties that are essentially continuous or graded in nature. The representation of two different sounds tends to overlap, with the degree of overlap dependent on where the inputs lie within the attractor structure encoded in the weights in the network. When the overlap is initially very high, adaptive training proceeds slowly and fixed training does not progress at all. For those subjects for whom the overlap is initially lower, adaptive training proceeds very quickly and there is also progress in the fixed training condition.

Incorporating a role for outcome information into Hebbian models of learning

While the results thus far suggest that we can go some distance in understanding the conditions under which human subjects succeed and fail to learn, relying only on the principle of Hebbian learning, it is also crucial to consider how other factors might shape the learning process. There are many studies in the literature indicating that mere exposure to stimuli that might elicit distinct percepts is not always enough to induce plasticity. An example from the extensive body of relevant work of Michael Merzenich and his associates will help illustrate this point. In one study (Recanzone et al, 1992a, Recanzone et al, 1992b) monkeys received vibratory stimulation when the surface of the middle finger of one hand was stimulated with a vibrating stylus. Monkeys who were required to discriminate different frequencies of vibration to obtain rewards showed improvement in their discrimination ability over training as well as dramatic reorganization of the sensory map representation for the stimulated hand, while monkeys who received yoked presentations of stimuli, but who were required to pay attention instead to other inputs, showed no reorganization.

Insert Figure 4 about here

With such results in mind, my collaborators and I wondered whether accuracy feedback might influence our Japanese listener's ability to learn the /r/-/l/ discrimination. To address this issue, we repeated our first experiment using fixed and adaptive training, but with the addition of visual feedback presented immediately after each response. Subjects received a row of three green checks if correct, or a row of three red x's if incorrect. The results of the experiment indeed confirmed that accuracy information can play a powerful role in enhancing learning: subjects in both the fixed and the adaptive group showed clear signs of learning (Figure 3, bottom panels), and both groups now showed clear signs of transfer of what they had learned to the untrained continuum. To our surprise, we found that with feedback, learning was fastest in the fixed, and not the adaptive training condition (Figure 4). Subjects in the fixed-with-feedback condition showed marked performance improvements at the outset of training, although there was some decline in performance between sessions, and showed very sharp identification functions after three days of training. These subjects also showed strong generalization to the untrained continuum and a sharp peak at the boundary between the /r/ and /l/ categories, consistent with the establishment of distinct categories or perceptual magnets for the /r/ and /l/ categories. Subjects in the adaptive-with-feedback condition did nearly as well, and overall slightly better than their counterparts in the adaptive-without-feedback conditions.

Integrating Outcome Information into Hebbian Learning Mechanisms

Our findings on the role of accuracy feedback provide one of many indications that there is something more to learning than the simple strengthening of synaptic connections among neurons that fire together. Yet our first experiment clearly indicates that outcome information is not always necessary for learning. This finding, together with the extensive support for Hebb-like learning mechanisms from LTP research, and the success of models using Hebbian mechanisms of learning in map-like neural structures to account for aspects of neural and cognitive development, suggests an important role for Hebbian learning mechanisms, or at least, some learning process that can operate to strengthen neural and behavioral responses in the absence of outcome information. Thus, my own inclination has been to consider ways in which a learning mechanism that operates according to Hebbian principles might be augmented by additional sources of information. Here I will briefly consider a few different possibilities. Note that this section is quite speculative in nature, but relates our work to that of others and raises points of general relevance to the possibility that a basically Hebbian mechanism is at work at the neural basis of learning and thus seems worth presenting in the present context.

One very simple proposal that can have dramatic effects within map-like representations is to modulate Hebbian learning as a function of attention. In fact, simply allowing attention to regulate neural activity per se provides one possible account for findings such as those of Recanzone et al above, in which there is massive reorganization of representations of attended, task-relevant stimuli. If the extent of neural activity itself is attention-dependent, and if (as many indications suggest) strong neural activity is needed to gate Hebbian plasticity, then the effects of task-related attention can be accommodated without any special regulation of learning based on outcome information per se.

While the above seems very likely to play some role in regulating the degree of learning, two specific points should be born clearly in mind. First, there are clearly signs of learning in experiments in which subjects are simply exposed to interesting stimuli, even when these are just presented in the background without any task (Saffran, Newport, and Aslin, 1996; Gomez, this volume). Of course it is not always clear in such studies whether subjects allocate covert attention to the stimuli, but the findings do suggest we need not assume that explicit task relevance and strong motivation to attend are always necessary for learning to occur. Second, it is not clear that a strictly attentional account provides a full basis for understanding the role of accuracy feedback within experiments where task demands are at least nominally held constant, as they are in the McCandliss et al experiments. All subjects in these experiments were attempting to learn to discriminate the /l/ and /r/ sounds we used. Accuracy feedback may well help to keep subjects motivated and thereby keep their attention focused, so we cannot rule out some role for a simple attentional effect, but it seems likely that some additional role for specific feedback is operative, over and above any such global effect.

There are several ways to use accuracy feedback to augment Hebbian learning. One possibility, based on O’Reilly’s LEABRA learning rule (O’Reilly, 1996; O’Reilly, this volume), is simply to combine error-correcting and Hebbian learning. In LEABRA, the signal that drives the connection weight combines the Hebb$_r$ learning rule of Equation 3 with an error-correcting term. If we were to apply this suggestion to capture the role of accuracy feedback in our /l/-/r/ learning experiment, we would need to imagine that the listener is able to translate the feedback signal into an indication of which response is correct, and then use this as the source of the correct target information required in standard error correcting learning, which then augments the Hebbian part of the learning when accuracy feedback is available.

The LEABRA approach is certainly worth exploring, but does introduce some processing complexity that has led me to consider other alternatives. To compute the error-correction component of the weight update, LEABRA uses a second pass through the activation settling process with the teaching input provided, after the first pass of activation in the absence of the teaching input. O’Reilly and I are currently at work on a successor to the LEABRA algorithm that attempts to eliminate the separate second pass. In the meantime, the two proposals considered below are perhaps mechanistically simpler than the existing version of LEABRA, and have thus been the focus of the modeling effort by Vallabha and McClelland (2004).

The first of the two ideas is to use the feedback signal to produce a reward signal I will call $R(F)$, and then use this to modulate Hebbian learning:

$$\Delta w_{rs} = \varepsilon R(F)\, f_r(a_r)\, f_s(a_s) \tag{4}$$

To apply this idea to the results of the experiments reviewed above, in which we see evidence of learning without any feedback, we would require that $R(F)$ have some positive value in the absence of any accuracy feedback. Feedback indicating that the response is correct could then increase the value of the $R(F)$ above its baseline value, and feedback indicating that the response is incorrect could reduce it below baseline, or potentially (as in many applications of reward-driven or reinforcement learning, c.f. Barto, 1992) reverse its sign. The second idea is to use the accuracy feedback signal to derive the identity of the correct response, and use this to adjust the activation of the output unit before applying Hebb$_r$ rule of

Equation 3. This approach treats the accuracy feedback as simply providing a source of activation to units representing the response alternatives.

In simulations, we have found that both of these approaches work fairly well to allow us to model the effects of accuracy feedback in the McCandliss et al (2002) experiments, and at present we have little basis for choosing among them. There are a few advantages to the reward-modulated learning approach, however, that may be worth mentioning. First, the mechanism requires only a global modulation of the extent of synaptic change, which from a neuro-mechanistic point of view is consistent with the fact that the reward signal comes with a slight delay after the convergent pre- and post-synaptic activity. Second, it is consistent with recent findings from a brain imaging study Tricomi, McClelland, and Fiez (2004) indicating that Japanese adults may treat accuracy feedback on their identification responses to /r/ and /l/ stimuli the same way subjects treat monetary reward (positive and negative monetary outcomes): Similar brain areas are activated with similar time courses for /r/-/l/ responses receiving positive feedback and for positive monetary reward provided for correct guesses in a random guessing task, and similar areas and time courses are also observed for incorrect /r/-/l/ responses and negative extrinsic reward in the random guessing task. The findings suggest that the subjects find accuracy feedback rewarding, and are consistent with the possibility that positive feedback gates the release of a global reward signal (e.g. dopamine) that could modulate synaptic plasticity (for a discussion of the role of dopamine as a reward signal, see Tricomi et al, 2004).

Summary and Conclusion

I have considered the idea of approaching human (and neural network) learning from an essentially Hebbian perspective. I have reviewed evidence from LTP studies and considered a range of computational models that incorporate Hebbian learning, and I have described experimental studies consistent with the idea, which seems to follow from a Hebbian approach, that there are processes at work strengthening the response a person makes to an input, either when this is incorrect, as in the apparent undesirable strengthening of incorrect associative responses in amnesia, or when there is no feedback, as in the adaptive-no-feedback condition of our experiment teaching Japanese adults to differentiate between /l/ and /r/. Hebbian learning can provide a basis for thinking about how we may learn from our own responses to things, in the absence of external teaching information, and such a process may also explain many cases of failures of learning, if circumstances conspire such that incorrect or unhelpful responses are elicited instead of correct or helpful ones. Based on these points, it seems likely that it will be worthwhile to consider the possible role of a Hebbian learning mechanisms in other cases of successes and failures of learning and development besides those considered in the present article.

Clearly, though, Hebb's proposal by itself is insufficient to address all aspects of human learning and memory. If such a process operates at all, it must be subject to extensive regulation and/or supplementation with other processes. Furthermore, the process operates within an organized network of interacting brain structures that play important roles in guiding processing toward correct outcomes. More work is necessary to understand these regulatory and/or supplemental processes, and to fully understand how interactions among brain regions, particularly medial temporal structures and neocortical learning systems, work together to achieve a system that functions successfully as an integrated whole.

References

Abbott, L. F., & Nelson, S. B. (2000). Synaptic plasticity: Taming the beast. Nature Neuroscience, 3, 1178–1183.

Baddeley, A., & Wilson, B. A. (1994). When implicit learning fails: Amnesia and the problem of error elimination. Neuropsychologia, 32, 53-68.

Barrionuevo, G., & Brown, T. H. (1983). Associative long-term synaptic potentiation in hippocampal slices. Proceedings of the National Academy of Science, USA, 80, 7347-7351.

Barto, A. G. (1992). Reinforcement learning and adaptive critic methods. In D. A. White & D. A. Sofge (Eds.), Handbook of intelligent control: Neural, fuzzy, and adaptive approaches (p. 469-491). New York: Van Nostrand Reinhold.

Bienenstock, E. L., Cooper, L. N., & Munro, P. W. (1982). Theory of the development of neuron selectivity: orientation specificity and binocular interaction in visual cortex. Journal of Neuroscience, 2, 32–48.

Bird, C., Cipolotti, L., Kwok, K., & McClelland, J. L. (2004). Successes and failures of associative learning in amnesia. Manuscript, Center for the Neural Basis of Cognition, Carnegie Mellon.

Cipolotti, L., Shallice, T., Chan, D., Fox, N., Scahill, R., Harrison, G., Stevens, J., & Rudge, P. (2001). Long-term retrograde amnesia...the crucial role of the hippocampus. Neuropsychologia, 39, 151-172.

Cohen, L. B., Chaput, H. H., & Cashon, C. H. (2002). A constructivist model of infant cognition. Cognitive Development, 17, 1323–1343.

Crowley, J. C., & Katz, L. C. (2000). Early development of ocular dominance columns. Science, 290, 1321–1324.

Dudek, S. M., & Bear, M. F. (1993). Bidirectional long-term modification of synaptic effectiveness in the adult and immature hippocampus. Journal of Neuroscience, 12(7), 2910–1918.

Elman, J. L. (1990). Finding structure in time. Cognitive Science, 14, 179-211.

Gabrieli, J. D. E., Cohen, N. J., & Corkin, S. (1988). The impaired learning of semantic knowledge following bilateral medial temporal-lobe resection. Brain and Cognition, 7, 157-177.

Gomez, R. (2005). Dynamically guided learning. In Y. Munakata & M. H. Johnson (Eds.), Processes of change in brain and cognitive development: Attention and performance XXI. Oxford: Oxford University Press.

Goodhill, G. J. (1993). Topography and ocular dominance: a model exploring positive correlations. Biological Cybernetics, 109-118.

Grossberg, S. (1976). On the development of feature detectors in the visual cortex with applications to learning and reaction-diffusion systems. Biologcial Cybernetics, 21, 145-159.

Hamman, S. B., & Squire, L. R. (1995). On the acquisition of new declarative knowledge in amnesia. Behavioral Neuroscience, 109, 1027-1044.

Hayman, C. A. G., MacDonald, C. A., & Tulving, E. (1993). The role of repetition and associative interference in new semantic learning in amnesia: A case experiment. Journal of Cognitive Neuroscience, 5, 375-389.

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

Inhelder, B., & Piaget, J. (1958). The growth of logical thinking from childhood to adolescence. New York: Basic Books.

Johnson, J., & Newport, E. (1989). Critical period effects in second-language learning: The influence of maturational state on the acquisition of english as a second language. Cognitive Psychology, 21, 60-99.

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

Kwok, K. (2003). A computational investigation into the successes and failures of semantic learning in normal humans and amnesics. Unpublished doctoral dissertation, Carnegie Mellon University, Department of Psychology.

Levy, W. B., & Steward, O. (1979). Synapses as associative memory elements in the hippocampal formation. Brain Research, 175, 233-245.

Linsker, R. (1986a). From basic network principles to neural architecture, III: Emergence of orientation columns. Proceedings of the National Academy of Sciences, USA, 83, 8779-8783.

Linsker, R. (1986b). From basic network principles to neural architecture, II: Emergence of orientation-selective cells. Proceedings of the National Academy of Sciences, USA, 83, 8390-8394.

Linsker, R. (1986c). From basic network principles to neural architecture, I: Emergence of spatial-opponent cells. Proceedings of the National Academy of Sciences, USA, 83, 7508-7512.

Margolis, H. (1987). Patterns, thinking, and cognition. Chicago, IL: University of Chicago Press.

Margolis, H. (1993). Paradigms and barriers. Chicago, IL: University of Chicago Press.

McCandliss, B. D., Fiez, J. A., Protopapas, A., Conway, M., & McClelland, J. L. (2002). Success and failure in teaching the [r]-[l] contrast to Japanese adults: Tests of a Hebbian model of plasticity and stabilization in spoken language perception. Cognitive, Affective & Behavioral Neuroscience, 2(2), 89-108.

McClelland, J. L. (1989). Parallel distributed processing: Implications for cognition and development. In R. G. M. Morris (Ed.), Parallel distribured processing: Implications for psychology and neurobiology (p. 8-45). New York: Oxford University Press.

McClelland, J. L. (1994). The interaction of nature and nurture in development: A parallel distributed processing perspective. In P. Bertelson, P. Eelen, & G. D'Ydewalle (Eds.), International perspectives on psychological science, Volume 1: Leading themes (p. 57-88). Hillsdale, NJ: Erlbaum.

McClelland, J. L. (1995). A connectionist perspective on knowledge and development. In T. J.

Simon & G. S. Halford (Eds.), Developing cognitive competence: New approaches to process modeling (p. 157-204). Hillsdale, NJ: Erlbaum.

McClelland, J. L. (2001). Failures to learn and their remediation: A Hebbian approach. In J. L. McClelland & R. S. Siegler (Eds.), Mechanisms of cognitive development: Behavioral and neural perspectives (p. 97-121). Mahwah, NJ: Erlbaum.

McClelland, J. L., & Jenkins, E. (1991). Nature, nurture, and connections: Implications of connectionist models for cognitive development. In K. V. Lehn (Ed.), Architectures for intelligence (p. 41-73). Hillsdale, NJ: Erlbaum.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. Psychological Review, 102, 419-457.

McClelland, J. L., Thomas, A., McCandliss, B. D., & Fiez, J. A. (1999). Understanding failures of learning: Hebbian learning, competition for representational space, and some preliminary experimental data. In J. A. Reggia, E. Ruppin, & D. Glanzman (Eds.), Progress in brain research (Vol. 121, p. 75-80). Amsterdam: Elsevier Science.

McNaughton, B. L., Douglas, R. M., & Goddard, G. V. (1978). Synaptic enhancement in facia dentata: Cooperativity among coactive afferents. Brain Research, 157, 277-293.

Miller, K. D., Keller, J. B., & Stryker, M. P. (1989). Ocular dominance column development: Analysis and simulation. Science, 245, 605-615.

Milner, B., Corkin, S., & Teuber, H.-L. (1968). Further analysis of the hippocampal amnesia syndrome: 14-year follow-up study of H.M. Neuropsychologia, 6, 215-234.

Munakata, Y., McClelland, J. L., Johnson, M. H., & Siegler, R. (1997). Rethinking infant knowledge: Toward an adaptive process accout of successes and failures in object permanence tasks. Psychological Review, 104, 686-713.

O'Reilly, R. (1996). The LEABRA model of neural interactions and learning in the neocortex. Unpublished doctoral dissertation, Department of Psychology, Carnegie Mellon University, Pittsburgh, PA.

O'Reilly, R. C. (2005). Modeling integration and dissociation in brain and cognitive development. In Y. Munakata & M. H. Johnson (Eds.), Processes of change in brain and cognitive development: Attention and performance XXI. Oxford: Oxford University Press.

Recanzone, G. H., Merzenich, M. M., Jenkins, W. M., Grajski, K. A., & Dinse, H. R. (1992). Topographic reorganization of the hand representation in cortical area 3b of owl monkeys trained in a frequency-discrimination task. Journal of Neurophysiology, 67, 1031-1056.

Recanzone, G. H., Merzenich, M. M., & Schreiner, C. E. (1992). Changes in the distributed temporal response properties of SI cortical neurons reflect improvements in performance on a temporally-based tactile discrimination task. Journal of Neurophysiology, 67, 1071-1091.

Rogers, T. T., & McClelland, J. L. (2004). Semantic cognition: A parallel distributed processing approach. Cambridge, MA: MIT Press.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. Nature, 323(9), 533-536.

Rumelhart, D. E., & Zipser, D. (1985). Feature discovery by competitive learning. Cognitive Science, 9, 75-112.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-olds. Science, 274(5294), 1926-1928.

Schyns, P. G. (1991). A modular neural network model of concept acquisition. Cognitive Science, 15, 461-508.

Scoville, W. B., & Milner, B. (1957). Loss of recent memory after bilateral hippocampal lesions. Journal of Neurology, Neurosurgery, and Psychiatry, 20, 11-21.

Seidenberg, M. S., & Zevin, J. D. (2005). Connectionist models in developmental cognitive neuroscience: Insights about critical periods. In Y. Munakata & M. H. Johnson (Eds.), Processes of change in brain and cognitive development: Attention and performance XXI. Oxford: Oxford University Press.

Siegler, R. S. (1976). Three aspects of cognitive development. Cognitive Psychology, 8(4), 481-520.

Song, S., Miller, K. D., & Abbott, L. F. (2000). Competitive Hebbian learning through spike-timing-dependent synaptic plasticity. Nature Neuroscience, 3, 919–926.

St. John, M. F., & McClelland, J. L. (1990). Learning and applying contextual constraints in sentence comprehension. Artificial Intelligence, 46, 217-257.

Tricomi, E. M., Delgado, M. R., & Fiez, J. A. (2004). Modulation of caudate activity by action contingency. Neuron, 41, 281–292.

Tricomi, E. M., McClelland, J. L., & Fiez, J. A. (2004). Japanese adults show reward-like brain activity when receiving feedback on correct responses during training to discriminate English /r/ and /l/. Manuscript, Center for the Neural Basis of Cognition, University of Pittsburgh and Carnegie Mellon.

Vallabha, G. K., & McClelland, J. L. (2004). A hebbian model of speech perceptual learning. Manuscript, Center for the Neural Basis of Cognition, Carnegie Mellon.

Author Note

Preparation of this article was supported by MH 64445 from the National Institute of Mental Health (USA). Correspondence regarding this article may be sent to James L. McClelland (jlm@cnbc.cmu.edu), CNBC, 115 Mellon Institute, 4400 Forbes Avenue, Pittsburgh, PA, 15213

Table 1. Effect of input correlation on connection weights.

event	a_1	a_2	a_3	a_r	$a_r a_1$	$a_r a_2$	$a_r a_s$
a	1	1	0	.20	.20	.20	0
b	0	0	1	.10	0	0	.10
sum of co-products					.20	.20	.10
decay					-.15	-.15	-.15
net change to weight					.05	.05	-.05

Effect of correlation of input activations on Hebbian learning. Two events are illustrated, one in which input units 1 and 2 are both active, and one in which input unit 3 is active alone. In the first case, the resulting activation of the receiving unit is greater than in the second case, so the Hebbian co-products ($a_r a_s$) are larger in the first case than the second case. For this case we are assuming $a_r = \Sigma_s a_s w_{rs}$ and $w_{rs} = .1$ for all s before either event is presented. Note that the co-products are all positive. The value chosen for weight decay determines how strong the sum of the co-products must be for the change in weight to be positive. Note also that changes in the weights will influence the activation of the receiving unit if the inputs are presented again; this tends to amplify the effect of the correlation in the inputs still further. Based on Linsker (1996a).

Figure Captions

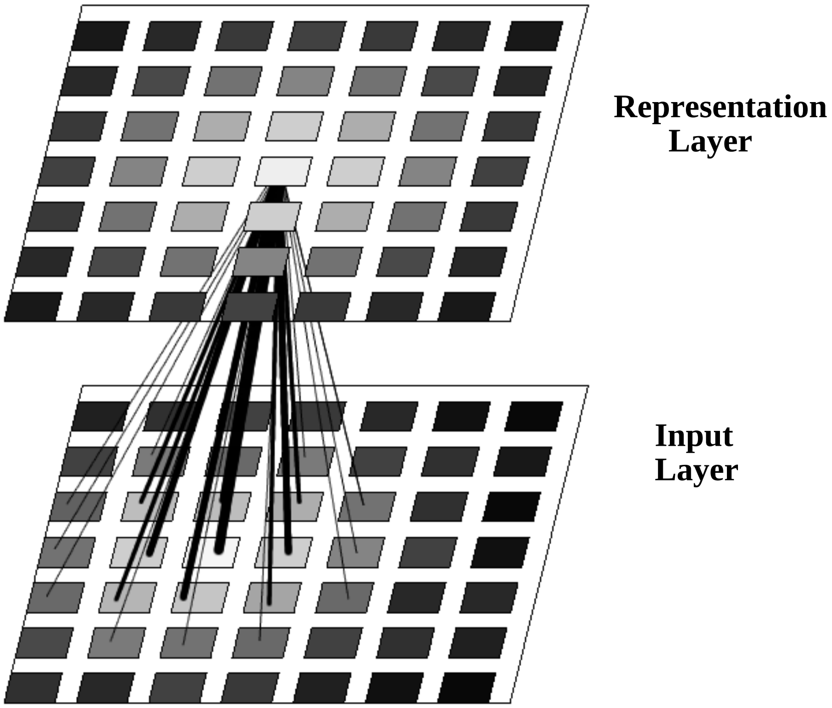
Figure 1. A simple map-formation network similar to the architecture proposed by Kohonen (1982). Units are represented by tiles within each layer, and the activation of each unit is indicated by shading (white = 1.0; black = 0). The network is shown with a blob of activation centered just to the left of the middle of the input layer. This in turn has produced a blob of activation centered on the middle unit of the representation layer. Connections from the center unit on the representation layer to the units on the input layer are illustrated, with thicker lines indicating stronger connections. These connections were established through learning; before learning, the connections were largely random, with only a weak tendency for connections from a particular input unit to project more strongly to the corresponding unit in the representation layer. Reprinted from Figure 1 of McClelland et al. (1999), permission pending.

Figure 2. Mean categorization functions of 12 native English speakers for synthesized speech stimuli from each of the two continua used in the experiments. The X axis represents the position on the stimulus in relation to the anchor stimuli, which are resynthesized versions of naturally spoken stimuli without exaggeration or interpolation. Percentages of trials eliciting R responses are plotted on the Y axis for each stimulus. Large empty circles represent the anchor stimuli resynthesized from the recorded base stimuli. Data points between the anchor stimuli are responses to stimuli interpolated between these anchors, and data points in the peripheral regions represent responses to extrapolated speech stimuli. Stimuli used for the fixed training condition are indicated with large filled circles. Triangles point to the positions of the initial stimuli used in the adaptive training condition. Reprinted from Figure 1, p. 92 in B.D McCandliss, J.A. Fiez, A. Protopapas, M. Conway, & J. L. McClelland, Success and failure in teaching the /r/-/l/ contrast to Japanese adults: Test of a Hebbian model of plasticity and stabilization in spoken language perception. Cognitive Affective, and Behavioral Neuroscience, 2002, 2, 89-108. Permission Pending.

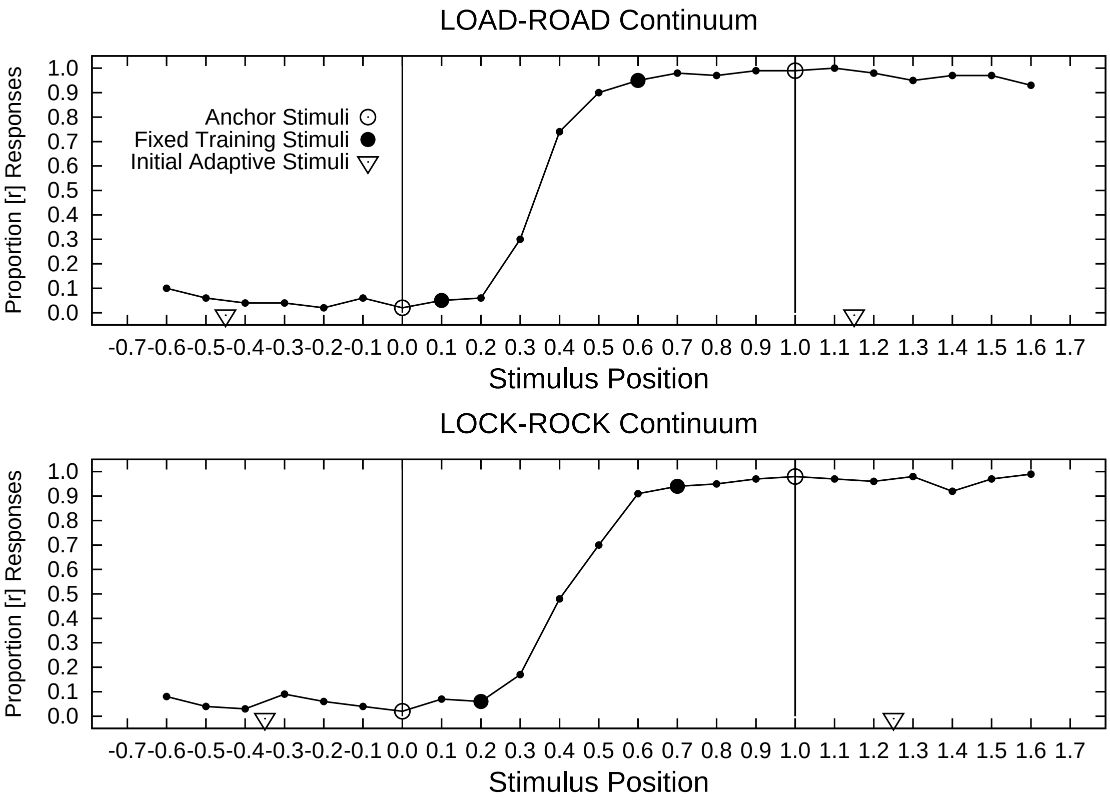
Figure 3. Mean categorization functions (with standard error bars) for four groups of Japanese subjects (n=8) before and after three twenty-minute training sessions in the four training conditions of the experiment. Pre- and post-test results are shown on the continuum used in training and on the other continuum used to assess transfer. Adapted from Figure 2, p. 95, and Figure 6, p. 99, in B.D McCandliss,

J.A. Fiez, A. Protopapas, M. Conway, & J. L. McClelland, Success and failure in teaching the /r/-/l/ contrast to Japanese adults: Test of a Hebbian model of plasticity and stabilization in spoken language perception. Cognitive Affective, and Behavioral Neuroscience, 2002, 2, 89-108. Permission Pending.

Figure 4. Mean percent correct on probe trials for subjects in each of the four training conditions over the course of training. Each data point is based on 10 probe trials per subject and encompasses 100 training trials. Data from two subjects in the fixed/no feedback condition have been excluded due to a data recording error on day 3. Their data were typical of the group on days 1 and 2. Reprinted from Figure 9, p. 102, in B.D McCandliss, J.A. Fiez, A. Protopapas, M. Conway, & J. L. McClelland, Success and failure in teaching the /r/-/l/ contrast to Japanese adults: Test of a Hebbian model of plasticity and stabilization in spoken language perception. Cognitive Affective, and Behavioral Neuroscience, 2002, 2, 89-108. Permission Pending.


Representation
Layer
Input
Layer

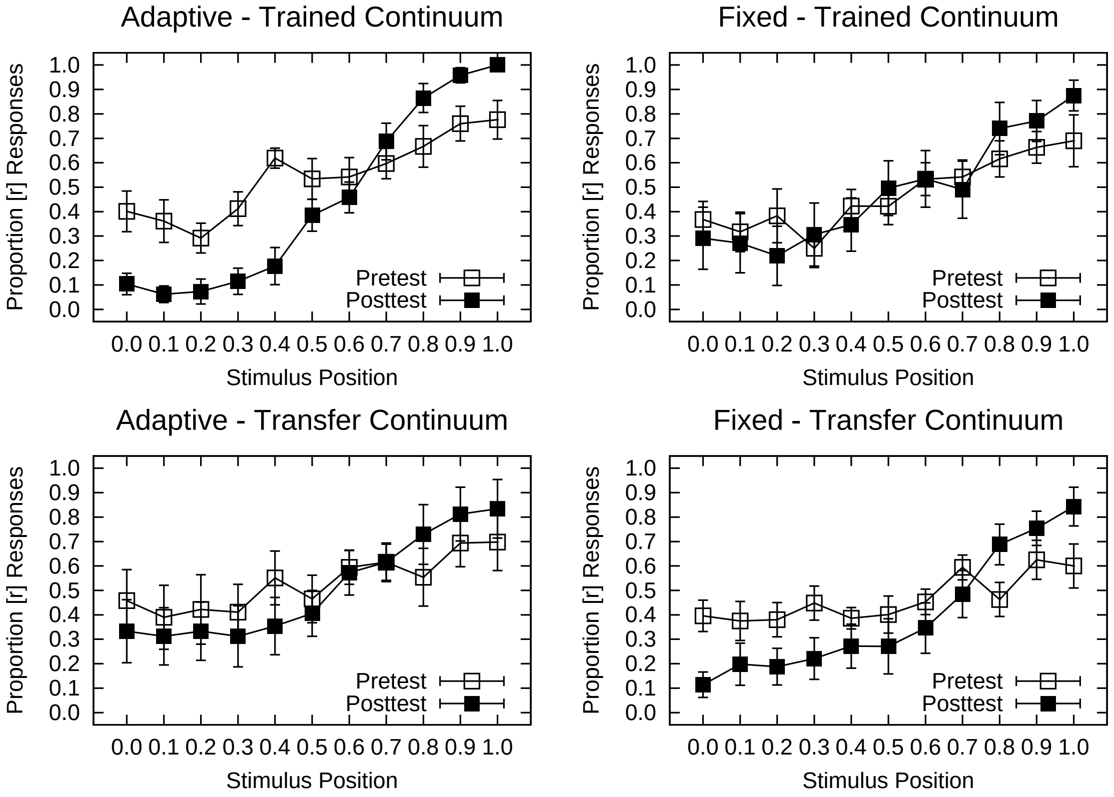
Native Speaker Identification Functions



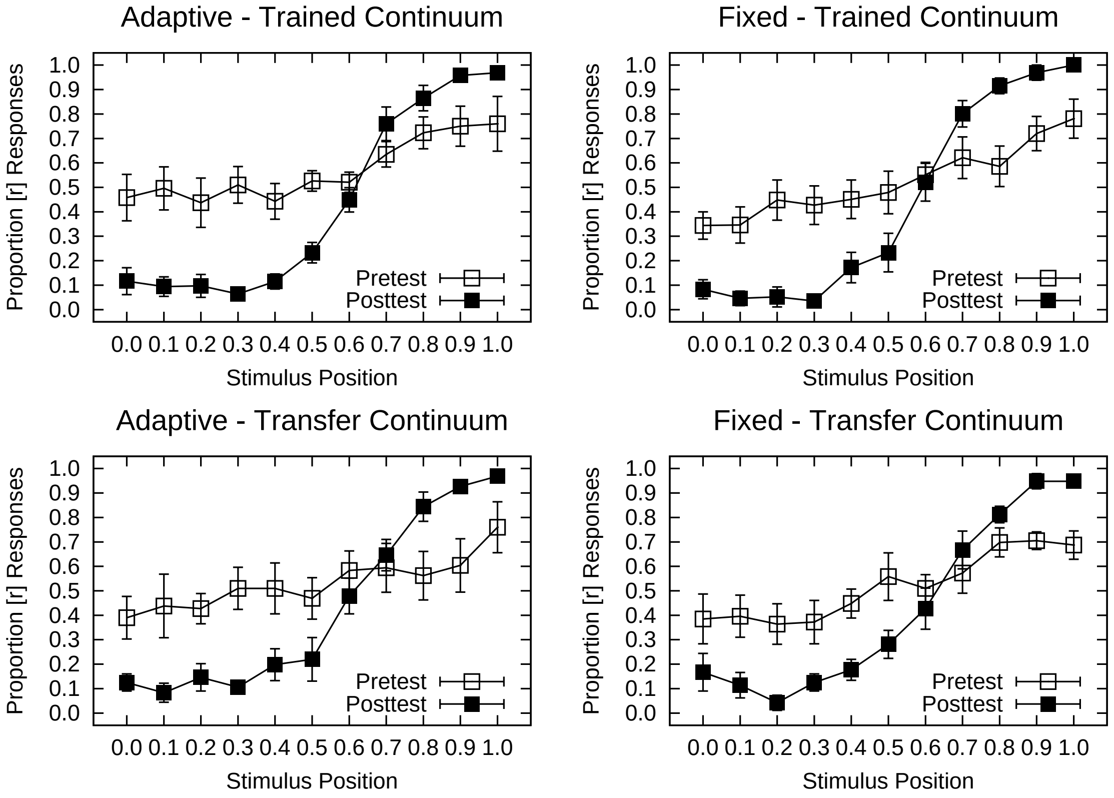
LOAD-ROAD Continuum

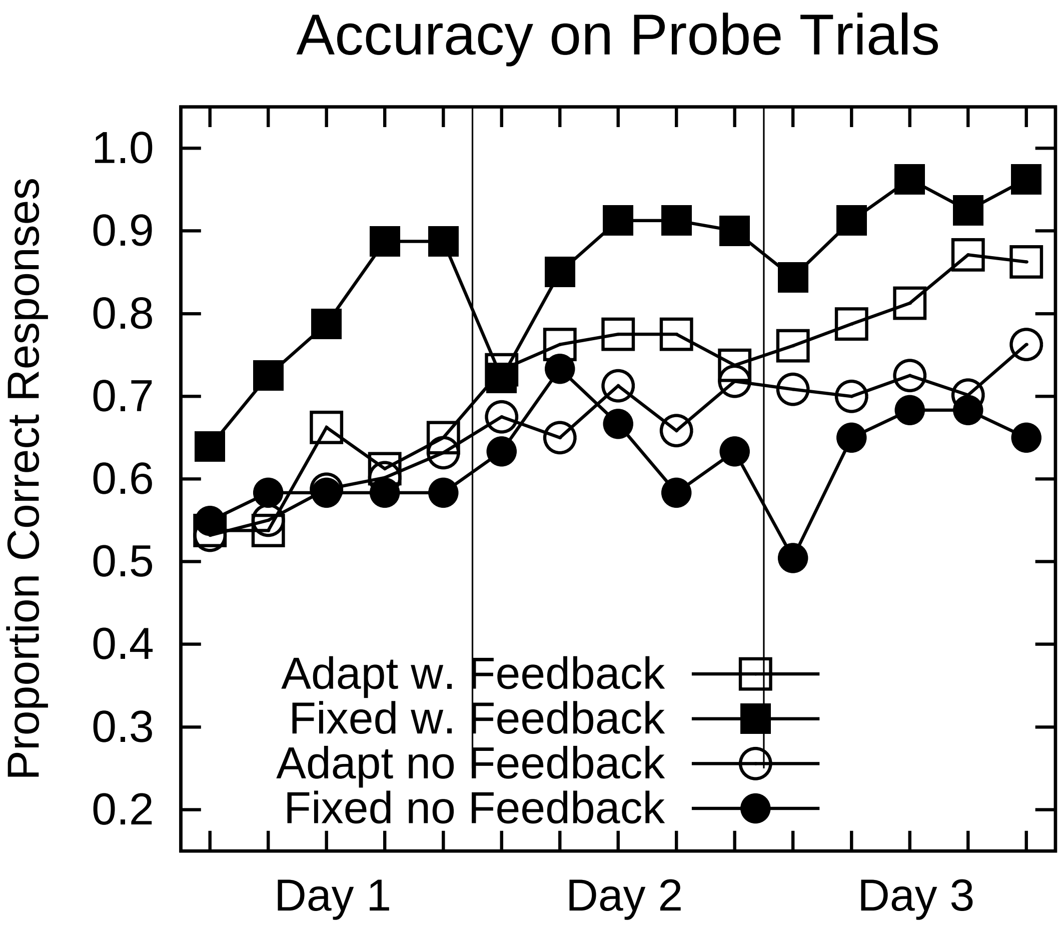
LOCK-ROCK Continuum

Effects of Training Without Feedback



Effects of Training With Feedback




Accuracy on Probe Trials
Proportion Correct Responses
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
Adapt w. Feedback □
Fixed w. Feedback ■
Adapt no Feedback ○
Fixed no Feedback ●
Day 1
Day 2
Day 3