In K. Van Lehn (1991). Architectures for Intelligence,
LEA; Hillsdale, NJ.

3 Nature, Nurture, and Connections: Implications of Connectionist Models for Cognitive Development

James L. McClelland
Eric Jenkins
Carnegie Mellon University

When it comes to selecting an architecture for modeling cognition, we have a choice. We can start with a symbolic architecture, in which the putative constituents of abstract cognition (symbols) are taken as modeling primatives; or we may adopt an alternative view, that symbolic *behavior* emerges from the operation of a system of simple, sub-symbolic processing units. Connectionist models take this latter tack. In these models, processing occurs through the propagation of activation among a number of simple processing units. The knowledge that governs processing is stored in the strengths of the connections among the units. And learning occurs through the gradual adjustment of the strengths of these connections. At first glance it may seem that such mechanisms are far removed from symbolic thought. Yet we will argue in this chapter that they may form the basis of the acquisition of a number of cognitive abilities, and that they may help us answer basic questions about the process of cognitive development. Several different kinds of answers have been given to these questions. We will see how the connectionist framework opens them anew and suggests what may be different answers in many cases.

THE PHENOMENA

The field of cognitive development is replete with examples of dramatic changes in children's thinking as they grow older. Here we give three examples: (a) Failures of conservation and compensation, (b) Progressive differentiation of knowledge about different kinds of things, (c) U-shaped learning curves in language acquisition.

Failures of Conservation and Compensation

Perhaps the best known phenomena in cognitive development are the dramatic failures of conservation that Piaget has reported in a wide range of different domains. One domain is the domain of liquid quantity. A child of 3 is shown two glasses of water. The glasses are the same, and each contains the same amount of water, and the child sees that the amount is the same. But when the contents of one of the glasses is poured into a wider container, the child will say that there is less liquid in the wider container.

It is typical to say that this answer that the young child gives reflects a failure to recognize two things: (a) That quantity is conserved under the transformation of pouring from one container to another; and (b) that greater width can compensate for less height. Many tasks are specifically designed to tap into the child's ability to cope with these kinds of compensation relations between variables.

One such task developed by Inhelder and Piaget (1958), the so-called *balance-scale task*, is illustrated in Fig. 3.1. In this task, the child is shown a balance scale with pegs at evenly spaced intervals to the left and right of a fulcrum. On one peg on the left are several weights; on one peg on the right are several weights. The scale is immobilized, and the child is asked to judge which side will go down, or whether they will balance. We will have occasion to examine performance in this task at length below; for now it suffices to note that young children (up to about 6 or 7 in this case) typically respond as if the distance from the fulcrum was completely irrelevant. They will say the scale should balance if the weight is the same on both sides, regardless of distance. Otherwise they say the side with the greater weight will go down. These children, then, appear to miss the fact that lesser weight can be compensated for by greater distance. Typically by the age of 11 or so children have some appreciation for this trade off; the details of the developmental progression are quite interesting, as we shall see below.

Progressive Differentiation of Ontological Categories

Other researchers, studying different domains, have noticed other kinds of developmental progressions. Keil (1979) studied children's judgements about whether you could say things like "A rabbit is an hour long." He supposed such judgments tapped children's knowledge about different kinds of things. In these judgments, Keil was interested not in whether the child saw a sentence as true or false, but in whether the child felt that one could make certain kinds of predications (e.g., that something is an hour long) when the something is a member of a certain "ontological category" (e.g., living thing). Keil found that children were much more permissive in their acceptance of statements than adults were, but their permissiveness was not simply random. Rather, they would accept statements that over-extended predicates to categories near the ones they typically apply to, but would not extend them further. Thus some children will accept predications like "The rock is asleep," but not "The rock is an hour long." It was as though children's knowledge of what predicates apply to particular categories becomes progressively more and more differentiated, as illustrated in Fig. 3.2.

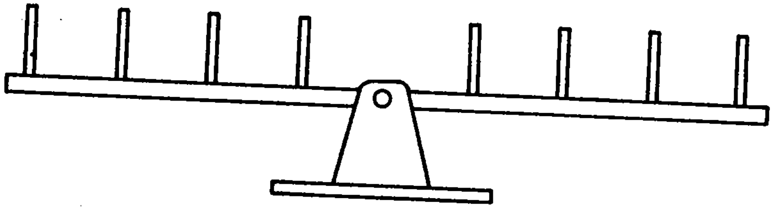


FIG. 3.1. Balance scale of the kind first used by Inhelder and Piaget (1958), and later used extensively by Siegler (1976; 1981; Siegler & Klahr, 1982). Reprinted from Siegler, 1976, Fig. 1, with permission.

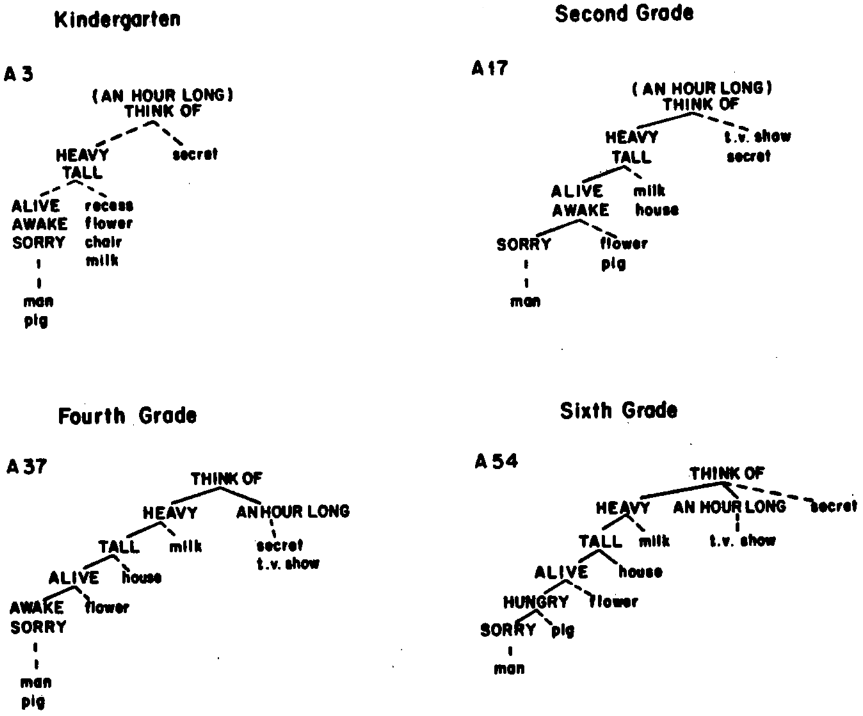


FIG. 3.2. Four different "predictability trees" illustrating the progressive differentiation of concepts as a function of age. Terms in capitals at internal nodes in the trees represent predicates, and terms in lower case at terminal nodes in the trees represent concepts that are spanned by all the predicates written on nodes that dominate the terminal. A predicate spans a concept if the child reports that it is not silly to apply either the predicate or its negation or both to the concept. Thus the first tree indicates that the child will accept "The girl is (not) alive," and "The chair is (not) tall" but will not accept "The chair is (not) alive." Parentheses indicate uncertainty about the application of a predicate. Redrawn from Keil, 1979, Appendix C, with permission.

U-Shaped Learning Curves in Language Development

Early on children often get certain kinds of linguistic constructions correct which they later get wrong; only later do they recover their former correct performance. One example is the passive construction, applied to semantically biased materials, such as "The man was bitten by the dog." (See Bever, 1970, for a discussion of the development of the use of the passive construction.) Early in development, children correctly interpret such sentences; they appear to be using information about what roles the different nouns typically play in the action described by the verb, since they tend to be correct only when the correct interpretation assigns the nouns to their typical roles. At an older age, children respond differently to such sentences, treating the first noun-phrase as the subject; semantic constraints are over-ridden, and there is a tendency to interpret "The man was bitten by the dog" as meaning "The man bit the dog." Finally, children interpret the sentence correctly again, but for a different reason. It would appear that they now know how to understand passives in general, since at this stage they can also interpret semantically neutral and even reverse-biased sentences (such as "the dog was bitten by the man") correctly.

THE QUESTIONS

The phenomena reviewed above raise basic questions about cognitive development. Three of these questions are:

- Are these different phenomena simply unrelated facts about development in different domains?
- Are there principles that all of these phenomena exemplify?
- If there are principles, are they domain specific, or are they general principles about development?

Different kinds of developmental theorists have answered such questions in very different ways. To Piaget, each failure of compensation or conservation reflected a single common developmental stage; the phenomena were intrinsically related by the characteristics of the stage, and these characteristics provided the basis for explanation.

Others have taken a very different approach. Keil (1979), following Chomsky's analogous argument for language, argued for *domain specific principles* of development. His view is that each cognitive domain has its own laws that provide constraints on what can be learned. These constraints limit the hypotheses that the child can entertain, thereby making it dramatically easier for the child to acquire adult abilities in the face of the impoverished information that is provided by experience with the world.

The main thrust of the remainder of this chapter is to argue that recent developments in connectionist learning procedures suggest a dramatic alternative to these kinds of views. The alternative is simply the hypothesis that these diverse developmental phenomena all reflect the operation of a single basic learning principle, operating in different tasks and different parts of the cognitive system.

THE LEARNING PRINCIPLE

The principle can be stated in fairly abstract terms as follows:

> Adjust the parameters of the mind in proportion to the extent to which their adjustment can produce a reduction in the discrepancy between expected and observed events.

This principle is not new. It might well be seen as capturing the residue of Piaget's accommodation process, in that accommodation involves an adjustment of mental structures in response to discrepancies. (See Flavell, 1963, for a discussion of Piaget's theory.) It is also very similar to the principle that governs learning in the Rescorla-Wagner model of classical conditioning (Rescorla & Wagner, 1972). What is new is that there exists a learning procedure for multilayer connectionist networks that implements this principle. Here, the parameters of the mind are the connections among the units in the network, and the procedure is the back propagation procedure of Rumelhart, Hinton, and Williams (1986).

The learning principle lies at the heart of a number of connectionist models that learn how to do various different kinds of information processing tasks, and that have applications to phenomena in cognitive and/or language development. Perhaps the simplest such model is the past-tense model of Rumelhart and McClelland (1986). The development of that model predated the discovery of the back propagation learning procedure, thereby forcing certain simplifications for the sake of developing an illustration of the basic point that lawful behavior might emerge from the application of a simple principle of learning to a connectionist network. Subsequent models have used back propagation to overcome some of these limitations. Included in this class are NETtalk (Sejnowski & Rosenberg, 1987) and a more recent model of word reading (Patterson, Seidenberg, & McClelland, 1989). The present effort grew out of two observations of similarities between the developmental courses seen in models embodying this principle, and the courses of development seen in children: First, the course of learning in a recent model of concept learning by Rumelhart (1990) is similar to aspects of the progressive differentiation of concepts reflected in Keil's (1979) studies of predictability. Second, the course of learning in a recent model of sentence comprehension by McClelland, St. John, and Taraban (1989) mirrors aspects of the progression from reliance on semantic constraints, to reliance on word order, to, finally, reliance on complex syntactic patterning such as the passive voice. We do not mean to claim that the models in question are fully adequate models of the developmental progression in either case; we only mean to claim that they seemed suggestive: They raised the possibility that part of the explanation of these and other developmental phenomena might be found in the operation of the learning principle as it adjusts connection strengths in a network subjected to patterns arising in its environment.

The remainder of this chapter presents two experiments assessing the applicability of this conjecture to another developmental phenomenon, namely the acquisition of the ability to take both weight and distance into account in the balance scale task described above. The task has been studied extensively by Siegler and his colleagues (Siegler, 1976, 1981; Siegler & Klahr, 1982), and quite a bit is known about it. We will first review the developmental findings. Then we will describe a connectionist model that captures these phenomena by applying the learning principle stated above (McClelland, 1989). As a follow-up, we will describe a second model that captures effects of specific experience on developmental change (Jenkins, 1986).

DEVELOPMENT OF JUDGMENTS OF BALANCE

In an important monograph, Siegler (1981) studied children's performance in the balance scale task and three other tasks in which two cues had to be taken into account for correct performance. In all cases, as in the balance scale task, the correct procedure requires multiplication. For example, in the balance scale task, to determine which side will go down, one must multiply the amount of weight on a given side of the beam times the distance of that weight from the fulcrum. The side with the greater product will go down; when the products are the same, the beam will balance.

Siegler studied children in several age groups, as well as young adults. Each child was asked to judge 24 balance problems. In each case, the scale was immobilized so that there was no feedback. The 24 problems could be divided into four of each six types:

- Balance. In this class of problem, the weight is the same on both sides of the scale and the weight is the same distance from the fulcrum on both sides.
- Weight. In these problems, the weights differ but distance from the fulcrum is the same on both sides.
- Distance. Here the weight is the same on both sides, but the distance from the fulcrum differs.

- Conflict. Here both weight and distance differ and are in conflict, in that the weight is greater on one side but the distance from the fulcrum is greater on the other. There are three types of conflict problems:
    - Conflict–weight. In these cases, the side with the greater weight has the greater torque (that is, the greater value of the product of weight times distance).
    - Conflict–distance. In these cases, the side with the greater distance has the greater torque.
    - Conflict–balance. Here the torques are the same on both sides.

Siegler's analysis of children's performance assumed that children use rule-governed procedures. Four such procedures or *rules* as Siegler called them are shown in Fig. 3.3. Each of these rules corresponds to a distinct pattern of performance over the six problem types. For example, children using Rule 1 should say the side with the greater weight will go down in weight problems and in all three types of conflict problems. They should think the scale will balance on balance problems and distance problems. In general, the mapping from the rules to expected performance is extremely straightforward. The only point that needs explication is the instruction *muddle through* when weight and distance conflict in Rule 3. In practice it is assumed to mean "guess randomly among the alternatives," so that 1/3 of the responses would be left-side-down; 1/3 right-side-down, and 1/3 balance.

Siegler compared the performance of each child tested with each rule, and counted discrepancies from predicted performance based on the rule. Children who scored less than four discrepancies from a given rule were scored as using that rule.

For our purposes, there are four basic findings that emerge from Siegler's analysis:

1. Lawful behavior. In general, performance of children over the age of 5 is extremely regular in the balance scale task. Overall about 90% of children tested conform to one of the four rules.

2. Developmental progression. As children get older, they appear to progress through the use of the different rules. The progression from Rule 1 to Rule 3 can be thought of as a progression in which at first the weight cue is relied on exclusively, while at the end distance and weight are both taken into account. In between (Rule 2), distance is taken into account only if it does not conflict with the weight cue. Children aged 5 to 7 typically use Rule 1, and college students typically use rules 3 or 4. Many college students do not have explicit knowledge of the torque principle. Children younger than age 5 tend not to be scorable strictly in terms of one of the rules; however, they appear to show an increasing tendency to behave in accordance with Rule 1.

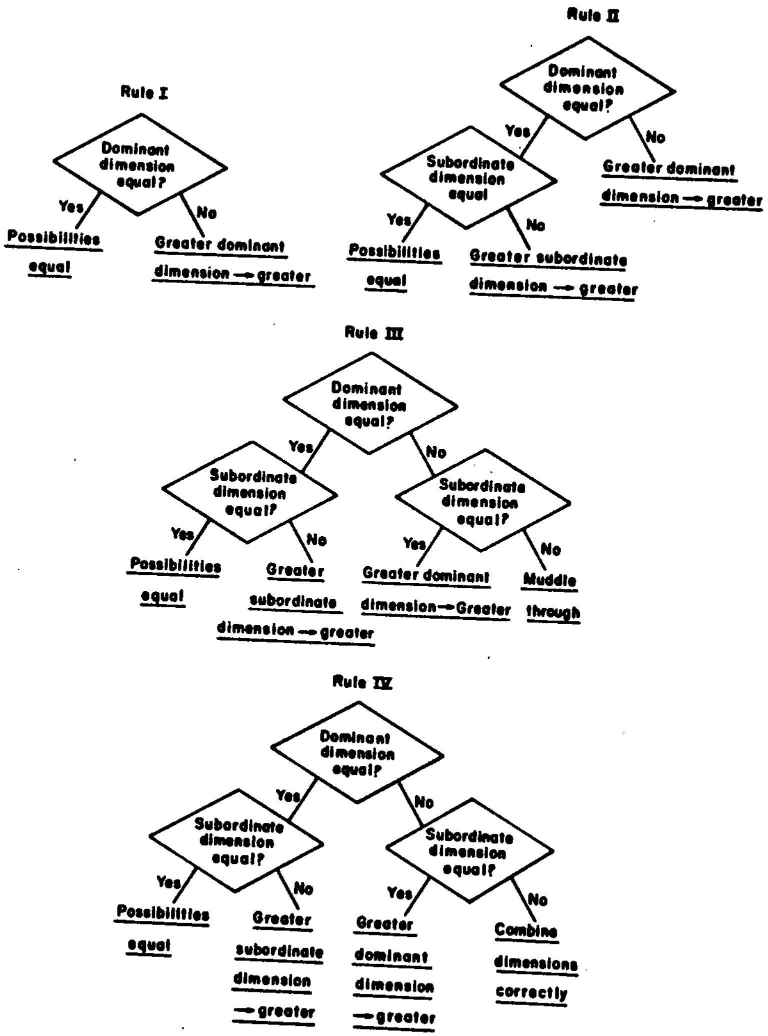

Rule I
Dominant dimension equal?
Yes
No
Possibilities equal
Greater dominant dimension → greater
Rule II
Dominant dimension equal?
Yes
No
Subordinate dimension equal
Greater dominant dimension → greater
Yes
No
Possibilities equal
Greater subordinate dimension → greater
Rule III
Dominant dimension equal?
Yes
No
Subordinate dimension equal?
Subordinate dimension equal?
Yes
No
Possibilities equal
Greater subordinate dimension → greater
Yes
No
Greater dominant dimension→Greater
Muddle through
Rule IV
Dominant dimension equal?
Yes
No
Subordinate dimension equal?
Subordinate dimension equal?
Yes
No
Possibilities equal
Greater subordinate dimension →greater
Yes
No
Greater dominant dimension →greater
Combine dimensions correctly

FIG. 3.3. Siegler's (1976, 1981) four "rules" for answering balance scale problems. Each Rule is in fact a full procedure, rather than a single rule. Reprinted with permission from Siegler (1981), Fig. 1.

3. Generality. The same four rules appear to be adequate to characterize performance in all three of the domains that Siegler studied. Though the developmental progression was not identical across domains, there was in all cases a trend from simpler to more complex rules with development.

4. Lack of correlation between domains. Even though children seem to progress through the same rules in different domains, they do not do so in lock-step; the correlation across domains is low, particularly in terms of the higher-numbered rules, so that children who are showing Rule 3 behavior in one task may be showing Rule 1 behavior in another and Rule 4 in a third.

MODEL OF THE BASIC PHENOMENA

The model we describe here was developed by McClelland (1989). It is based on earlier work by Jenkins (1986) relevant to other aspects of Siegler's data (Siegler, 1976; Siegler & Klahr, 1982) to which we will turn our attention below.

The model is sketched in Fig. 3.4. Of course, the model is a drastic oversimplification of the human mind and of the task; but as we shall see it allows us to capture the essence of Siegler's findings, and to see them emerge from the operation of the learning principle described above.

The model consists of a set of input units, to which balance problems can be presented as patterns of inputs; a set of output units over which the answer to

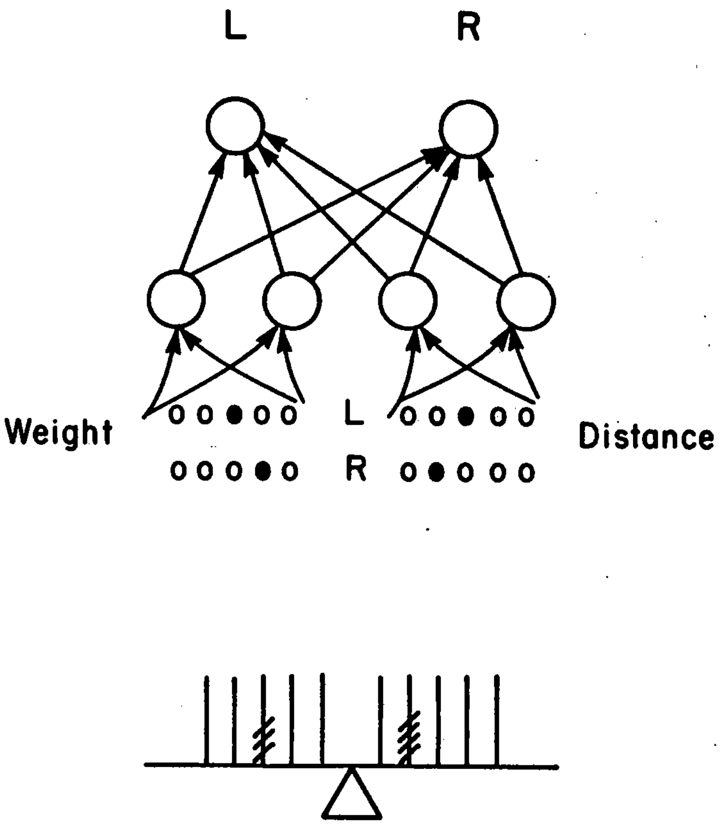


FIG. 3.4. The network used in the simulation of the development of performance in the balance scale task.

each problem can be represented; and a set of hidden units, between the input and the output. Connections run from input units to hidden units and from hidden units to output units.

The input units can be divided into two groups of 10. One group is used to represent information about weight and the other is used to represent information about distance. In each case the input representation imposes as little structure as possible on the input patterns. Each possible value of weight or distance from the fulcrum is assigned a separate unit. The ordering of values from low to high is not given in this representation; the network will have to learn this ordering. For the convenience of the reader, the units are arranged in rows according to which side of the beam they are from, and within each row they are arranged from left to right in order of increasing weight or distance from the fulcrum; but this ordering is unknown to the model before it is trained, as we shall see.

Though the two dimensions are not intrinsically structured for the model, the design of the network does impose a separate analysis of each dimension. This separation turns out to be critical; we will consider the implications of this architectural simplification below. The separation is implemented as follows: there are separate pairs of hidden units for each dimension. Two hidden units receive input from the weight input units and two receive input from the distance input units.

Each of the four hidden units projects to each of the two output units. The left output unit can be thought of as a "left side down" unit, and the right one as a "right side down unit." Thus a correct network for the task would turn on the output unit corresponding to the side with the greater torque, and would turn off the unit for the other side. For balance problems, we assume that the network should turn both units on half-way. Note that this coding of the output patterns does tell the network that balance is between left side down and right side-down.

Processing. Balance problems of the kind studied by Siegler can be processed by the network by simply turning on (i.e., setting to 1) the input units corresponding to a particular problem and turning off (i.e., setting to 0) all other input units. The input from the problem illustrated in Figure 3.7 is shown by using black to indicate those input units whose activations are 1.0, and white for the units whose activations are 0.

The inputs are propagated forward to the hidden units. Each hidden unit simply computes a net input:

$$net_i = \sum_j w_{ij}a_j + bias_i.$$

Here j ranges over the input units. Each hidden unit then sets its activation according to the logistic function:

$$a_i = \frac{1}{1 + e^{-net_i}}$$

In these equations, w_{ij} is the strength of the connection to hidden unit i from input unit j, a_j is the activation of input unit j, and $bias_i$ is the modifiable bias of hidden unit i. This bias is equivalent to a weight to unit i from a special unit that is always on.

Once activations of the hidden units are determined, the activations of the output units are determined by the same procedure. That is, the net input to each output unit is determined based on the activations of the hidden units, the weights from the hidden units to the output units, and the biases of the output units. Then the activations of the output units are determined using the logistic function.

Responses. The activations of the output units are real numbers between 0 and 1; to relate its performance to the balance scale task, these real-valued outputs must be translated into discrete responses. If the activation of one output unit exceeded the activation of the other by .333, the answer was taken to be "more active side down." Otherwise, the answer was assumed to be "both sides equal."

Learning. Before training begins, the strengths of these connections from input to hidden units and from hidden to output units are initialized to random values uniformly distributed between +.5 and −.5. In this state, inputs lead to random patterns of activity over both the hidden and output units. The activations of the output units fluctuate approximately randomly between about .4 and .6 for different input patterns. The network comes to respond correctly only as a result of training. Conceptually, training is thought of as occurring as a result of a series of experiences in which the network is shown a balance problem as input; computes activations of output patterns based on its existing connection weights; and is then shown the correct answer. The signal that drives learning is the difference between the obtained activation of each output unit and the correct or target activation for that unit. The back-propagation procedure of Rumelhart, Hinton, and Williams (1986) is then used to determine how each connection strength in the network should be adjusted to reduce these differences. Since the procedure is quite well-known, suffice it to say that it exactly implements the learning principle stated above, and restated here in network terminology:

> Adjust each weight in the network in proportion to the extent to which its adjustment can produce a reduction in the discrepancy between the expected event and the observed event, in the present context.

Here the "expected event" is the pattern of activation over the output units that is computed by the network, the observed event is the pattern of activation the environment indicates these units have, and the present context is the pattern of activation over the input units. Note that the direction of change to a connection (positive or negative) is simply the direction than tends to reduce the discrepancy between computed output and the correct or target output.

Environment. The environment in which a network learns plays a very strong role in determining what it learns, and particularly the developmental course of learning. The simulations reported here were based on the assumption that the environment for learning about balance problems consists of experiences that vary more frequently on the weight dimension than they do on the distance dimension. Of course, we do not mean to suggest that all the learning that children do that is relevant to their understanding of balance takes the form of explicit balance problems of the kind our network sees. Rather, our assumption that the experience on balance problems is dominated by problems in which there is no variability in weight is meant as a proxy for the more general assumption that children generally have more experiences with weight than with distance as a factor in determining the relative heaviness of something.[1]

The specific assumptions about the sequence of learning experiences were as follows. The environment consisted of a list of training examples containing the full set of 625 possible problems involving 25 combinations of possible weights (1 to 5 on the left crossed with 1 to 5 on the right) crossed with 25 combinations of possible distances (1 to 5 steps from the fulcrum on the left crossed with 1 to 5 steps from the fulcrum on the right). Two corpora were set up. Problems in which the distance from the fulcrum was the same on both sides were listed 5 times each in one corpus, and 10 times each in the other corpus. Other problems were listed only once in each corpus.

Training and testing regime. Four simulation runs were carried out, two with each of the two corpora just described. In each run, training consisted of a series of epochs. In each epoch, 100 patterns were chosen randomly from the full list of patterns in the corpus. In each epoch, weight increments were accumulated over the 100 training trials and then added into the weights at the end of the epoch, according to the momentum method described in Rumelhart, Hinton, and Williams (1986 p. 330); parameters were $\eta = 0.075$, $\alpha = .9$).

After weight updating at the end of each epoch, the network was given a 24 item test, containing four problems of each of the six types described above, taken from an experiment of Siegler's. (A few of the examples had to be modified since Siegler's experiment had used up to six pegs.)

A Comment on the Simulation Model

The model described above obviously simplifies the task that the learner faces and structures it for him to some degree. In particular, it embodies two principal assumptions which are crucial to the successful simulations we will consider below:

Environment Assumption. The model assumes that the environment is biased, so that one dimension—in this case weight—is more frequently available as a basis for predicting outcome than the other.

Architecture Assumption. The model assumes that the weight and distance dimensions are analyzed separately, before information about the two dimensions is combined.

Both of these assumptions are crucial to the success of the model. In an unbiased environment, both cues would be learned equally rapidly. Effects of combining the cues from the start as prescribed by the architecture assumption are more complex, but suffice it to say for now that the apparent stagelike character of performance is much less clear unless this assumption is adopted.

An important topic for further research will be to examine what variants of these assumptions might still allow the model to be successful. For example, regarding the environment, differences in salience (i.e., strength of input activations) and structuredness of the dimensions might also produce similar results.

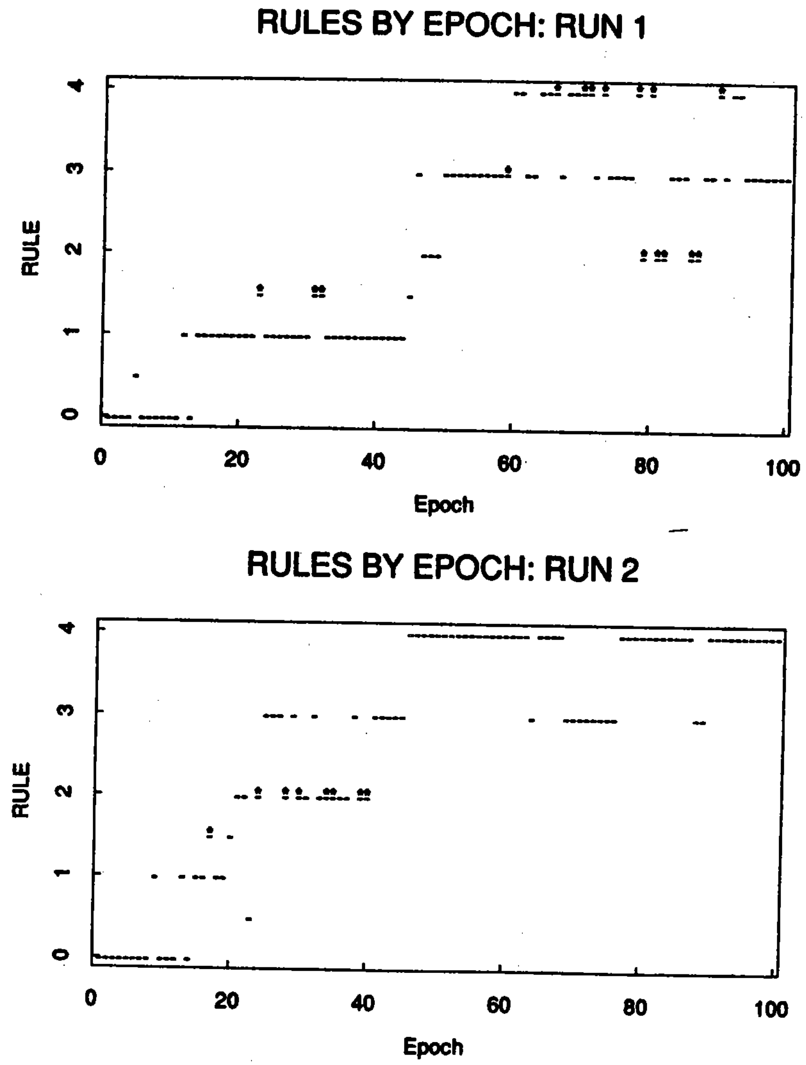
The issue of structuredness of the dimensions is a key point that needs to be considered as it relates to the present simulation. For both dimensions, the input representations encode different weights and distances from the fulcrum using distinct units. This means that different values are distinguishable by the model, but they are not structured for it; for example the input itself provides no indication that a distance or weight of 3 is between 2 and 4. The network must learn to represent the weights and distances in structured ways in order to solve the balance problem. We will see that it does this later.

Results

In general performance of the model conformed to one of the four rules described by Siegler. Over the four runs, the model fit the criteria of one of Siegler's four rules on 85% of the occasions, not counting an initial pre–Rule 1 period (In Siegler, 1981, the conformity figure is about 90%). Of course, the model was not consulting these rules or following the step-by-step procedures indicated in them; rather its behavior was simply scorable by Siegler's criteria as consistent with the succession of rules. Excluding the initial period, failures to fit the rules were of three types: Cases in which a rule fit except for a position bias that gave difficulty on balance problems, cases in which performance was borderline between Rules I and II, and combinations of these two problems. (Siegler [personal communication] does find some borderline cases between Rule 1 and Rule 2, but the position bias cases are not typical of children's performance.)

Overall Developmental Trends. Epoch by epoch performance in each of the four runs is shown in Figs. 3.5 and 3.6. One generally observes the expected developmental progression. Each simulation run is slightly different, due to differences in the random starting weights and the sequence of actual training

[1] An alternative assumption which might account for the developmental data just as well is the assumption that the weight dimension is pre-structured before the child comes to consider balance problems, while the distance dimension is not. The assumption that distance varies less frequently than weight but that neither dimension is initially structured allows us to observe the structuring process for both dimensions.


RULES BY EPOCH: RUN 1

RULES BY EPOCH: RUN 2

FIG. 3.5. Epoch by epoch performance of the simulation model in the two runs with a 5 to 1 bias favoring problems in which distance did not vary. Performance is scored by Rule. Cases marked by * missed a rule due to position bias. Rule 0 corresponds to always saying "balance," and occurs at the beginning of training. Rule 1.5 corresponds to performance on the borderline between Rules 1 and 2.

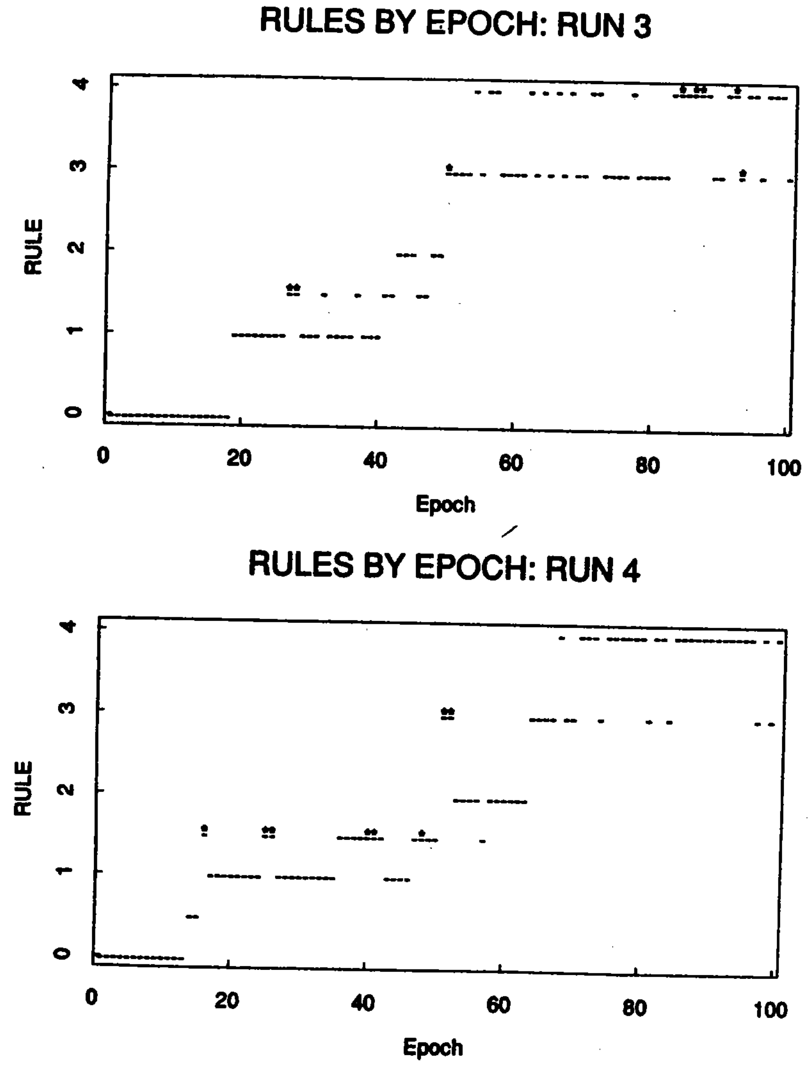
experiences, but there are clear common trends. Over the first 10 epochs or so, the output of the model was close to .5 on all test patterns; by our scoring criteria, all of these outputs count as "balance" responses, but of course they really represent a stage in which neither weight nor distance governs performance. The next few epochs represent a transition to Rule 1, in that in this phase the model is showing some tendency to activate the output unit on the side with the greater weight, but this tendency is variable across patterns and the discrepancy between the activations of the output units is not reliably greater than .33 when the weights differ.

After this brief transition, performance of the model has generally reached the point where it was responding consistently to the weight cue while systematically ignoring the distance dimension. This pattern continued for several more epochs. There was a brief transitional period, in which the model behave inconsistently on the *distance* problems crucial to distinguishing between Rule 1 and Rule 2 behavior. After several epochs in this phase, use of the distance cue reached the point where performance on all types of conflict problems became variable. The model generally continued in this phase indefinitely, sometimes reaching the point where its performance was generally scorable as fitting Rule 4 and sometimes not.

The variability in the model's performance from epoch to epoch is actually quite consistent with test-retest data reported in Siegler (1981). Rule 2 behavior is highly unstable, and there is some instability of behavior in other rules as well.

Performance in each phase. Siegler's criteria for conformity to his rules allow for some deviations from perfect conformity; in fact only 83% of test problems must be scorable as consistent with the rule. Given this, it is interesting to see whether the discrepancies from the rules that are exhibited by the model are consistent with human subject's performance. In general, they seem to be quite consistent, as Fig. 3.7 indicates. Each panel shows percent correct performance by the model averaged over the tests on which the model scored in accordance with one of the four rules. Also shown are data from two groups of human subjects as well as the pattern of performance that would be expected from a perfect rule user.

For Rule 1, the model differs very little from humans. For Rule 2, again the correspondence to human data is very close. Both the model and the humans show some slight tendency to get *conflict–distance* problems correct, and to occasionally miss *distance* and *balance* problems. For both Rule 1 and Rule 2, the tendency to miss *balance* problems is slightly greater in the model than in the children's data. For Rule 3, the model exaggerates a tendency seen in the human data to be correct on *conflict–weight* problems more often than on *conflict–distance* problems. The major discrepancy from the data is that the model is too accurate on conflict-balance problems. For Rule 4, the model again exaggerates a tendency seen in the human data to have residual difficulties with conflict problems.

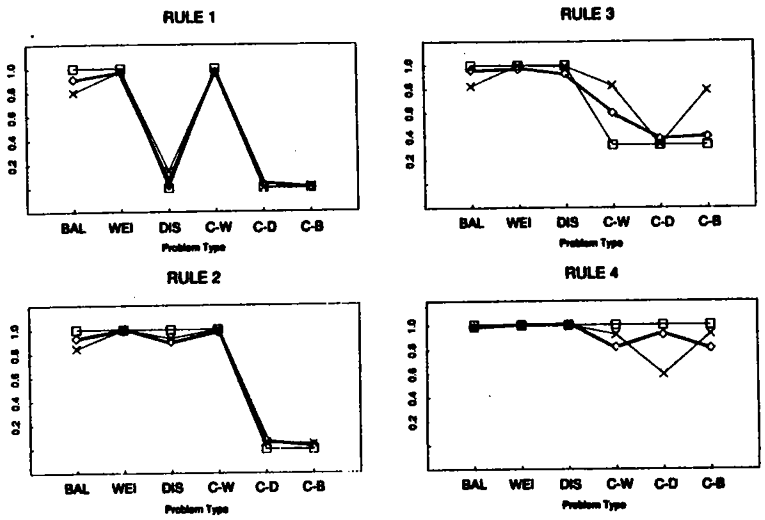


RULES BY EPOCH: RUN 3

RULES BY EPOCH: RUN 4

FIG. 3.6. Epoch by epoch performance of the simulation model in the two runs with a 10 to 1 bias favoring problems in which distance did not vary. Performance is scored by Rule, as in Fig. 3.8.

With the exception of the *conflict-balance* problems in Rule 3, the human data seem to fall about half-way between the model and perfect correspondence to the rules. It is tempting to speculate that some human subjects—particularly Rule 4 subjects—may in fact use explicit rules like the torque rule some of the time. It is, indeed, easy for the adult subjects who contribute to the Rule 4 results to



RULE 1

RULE 3

RULE 2

RULE 4

FIG. 3.7. Children's performance by problem type on the balance scale task, together with the performance of the simulation model and expected performance based on each rule. The heavy line with diamonds indicates children's performance. The model's performance is given by the light line with x's, whereas performance predicted from the rule is given by the light line with squares. For each child and each test of the simulation, performance was precategorized according to the best fitting rule. Then, percent correct responses by problem type were calculated averaging over children or simulation tests falling into each rule.

follow the torque rule if instructed specifically in this rule. However, it is evident that the subjects who fall under the Rule 4 scoring criteria do not in fact adhere exactly to the rule. Perhaps this group includes some individuals performing on the basis of implicit knowledge of the trade-off of weight and distance as well as some who explicitly use the torque rule, and perhaps some individuals use a mixture of the two strategies.

Further correspondences between the model and child development. So far we have seen that the balance scale model captures the pattern of development seen in the studies of Siegler (1976, 1981). There are two further aspects of the developmental data which are consistent with the gradual buildup of strength on the distance dimension that we see in the model:

1. Wilkening and Anderson (in press) present subjects with one side of a balance beam, and allow them to adjust the weight on the other side at a fixed distance from the fulcrum to make the scale balance. Over the age range of 9 to

20, in which children are generally progressing from late Rule 1 or Rule 2 to Rule 3 or Rule 4 according to Siegler's methods, they find an increasing sensitivity to the distance cue. Unfortunately it is difficult to be sure whether this reflects different numbers of subjects relying on the distance cue, or (as we see in the model) differences in degree of reliance among those who show some sensitivity to the distance cue.

2. For children who exhibit Rule 3 on Siegler's 24-item test, careful assessment with a larger number of conflict problems indicates the use of cue compensation strategies, rather than random guessing (Ferretti, Butterfield, Cahn, & Kerkman, 1985). Thus children are not simply totally confused about conflict problems during this stage but have some sensitivity of relative magnitudes of cues, as does the model. The exact degree of correspondence of the model's performance and human performance on these larger tests remains to be explored.

The mechanism for developmental change. Given the generally close correspondence between model and data, it is important to understand just how the model performs, and how its performance changes. To do this, it is helpful to examine the connections in the network at several different points in the learning process. Figure 3.8 displays the connections from the run that produced the results shown in the top panel of Fig. 3.6, at 4 different points during learning: At epoch 0, before any learning; at epoch 20, early in the Rule 1 phase; at epoch 40, at the end of the Rule 1 phase; and at epoch 100, when the simulation was terminated. Each of the four large rectangles in each panel shows the weights coming into and out of one of the four hidden units. The two on the left receive input from the weight dimension, and the two on the right receive input from the distance dimension.

In the first panel, before learning begins, all the connection strengths have small random values. In this situation, the output of the hidden units is not systematically related to magnitudes of the weights or distances, and is therefore of no use in predicting the correct output. At this point, the hidden units are not encoding either relative weight or relative distance, and are therefore providing no information that would be useful for predicting whether the left or right side should go down.

The first phase of learning consists of the gradual organization of the connections that process the amount of weight on each side of the balance scale. Recall that the network receives problems in which the distance cue varies much less frequently than problems in which the weight cue varies. Learning to rely on the weight cue proceeds more quickly than learning to rely on the distance cue as a simple result of this fact. The rate of learning with respect to each type of cue is relatively gradual at first, but then speeds up, for reasons that we will explore below. The relatively rapid transition from virtually unresponsive output to fairly strong reliance on the weight cue represents the brief transition to Rule 1 responding.

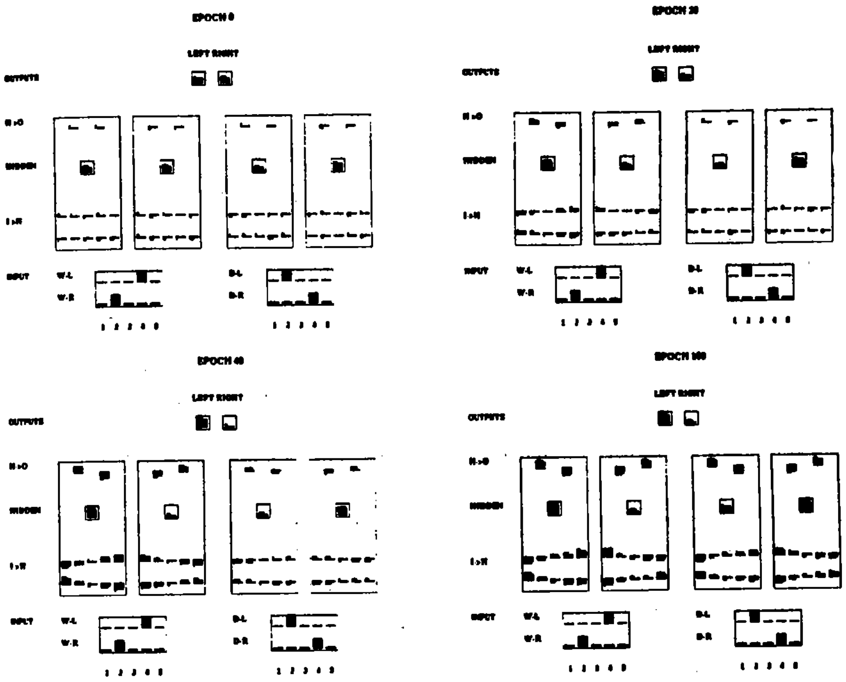


FIG. 3.8. Connection strengths into (I→H) and out of (H→O) each of the hidden units, at each of four different points during training. Activations of input, hidden, and output units are also shown, for a conflict balance problem, in which there are 2 weights on peg 4 on the left and 4 weights on peg 2 on the right. Magnitude of each connection is given by the size of the blackened area. Sign is indicated by whether the blackened area extends above or below the horizontal baseline. Note that activations are all positive, and range from 0 to 1. The connection strengths range between +6 and −6. See text for further explanation.

The result of this phase, in the second panel of the figure, is a set of connections that allow the hidden units on the left to reflect the relative amount of weight on the left vs. the right side of the balance scale. The leftmost hidden unit is most strongly excited by large weights on the left and small weights on the right, and most strongly inhibited by large weights on the right and small weights on the left. The activation of this unit, then, ranges from near 0 to near 1 as the relative magnitude of weight ranges from much more on the right to much more on the left. Correspondingly, this unit has an excitatory connection to the left-side-down output unit, and an inhibitory connection to the right-side-down output unit. The second hidden unit mirrors these relationships in reverse. At this point, then, the hidden units can be said to have learned to represent something they were not representing before, namely the relative magnitude of the inputs. Note that this information is not explicitly contained in the input, which simply distinguishes but does not order the different possible values of weight on the two sides of the balance scale.

At this point, the connection strengths in the distance part of the network remain virtually unchanged; thus, at the hidden unit level, the network has not yet learned to encode the distance dimension.

Over the next 20 epochs, connections get much stronger on the weight dimension, and we begin to see some organization of the distance dimension. While this is going on, the overt behavior of the network remains Rule 1 behavior. The network is getting ready for the relatively rapid transition to Rule 2 and then to Rule 3 which occurs over the next several epochs of training (as shown in the top panel of Fig. 3.6), but at epoch 40, the end of the Rule 1 phase, the distance connections are still not quite strong enough that they can yet push activations of the output units out of the balance range. With further learning, the distance cue becomes stronger and stronger; this first causes the distance cue to govern performance when the weights are in balance, giving rise to Rule 2 behavior. Further strengthening causes the distance cue to win out in some conflict problems, giving rise to behavior consistent with Rules 3 and 4. At epoch 100 of this particular run, the weight dimension maintains a slight ascendancy, so that with the particular *conflict–balance* problem illustrated, the model activates the left-side down unit, corresponding to the side with the greater weight, more than it activates the right-side down unit.

A couple of aspects of the developmental progression deserve comment. As Fig. 3.9 illustrates, the connection strengths are largely insensitive to differences early on, then go through a fairly rapid transition in sensitivity and then level off again. The acceleration seen in learning is a result of an inherent characteristic of the gradient descent learning procedure coupled with the architecture of the network. The procedure adjusts each connection in proportion to the magnitude of the effect that adjusting it will have on the discrepancy between correct and actual output. But the effect of a given connection depends on the strengths of other connections. Consider the connection coming into a hidden unit from one of the input units. An adjustment of the strength of this input connection will have a small effect on the output if the connections from the hidden unit to the output units are weak. In this case, the input connection will only receive a small adjustment. If however, the connections from the hidden units to the output units are strong, an adjustment of the strength of the input connection will have a much larger effect; consequently the learning procedure makes a much larger adjustment in this case. A slightly different story applies to the connections from the hidden units to the output units. When the connections from the input to the hidden units are weak and random, the activations of the hidden units are only weakly related to the correct output. Under these circumstances, the adjustments

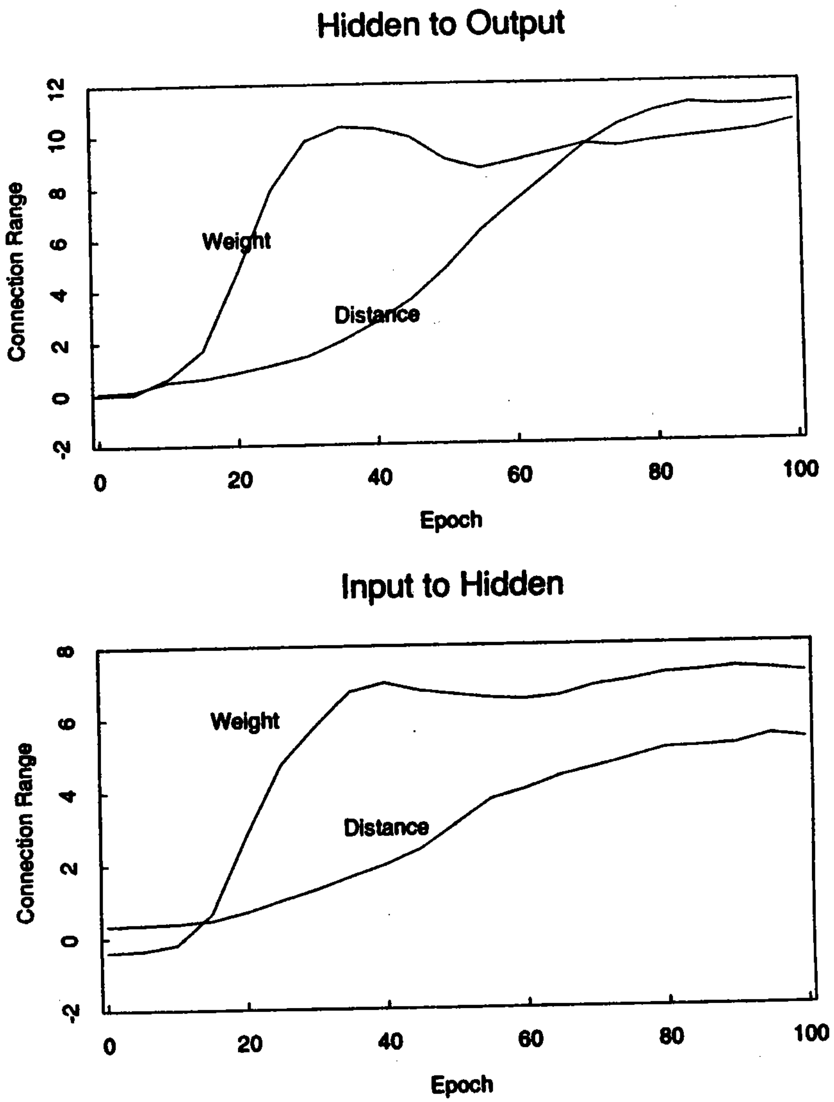


FIG. 3.9. Relative magnitude of connection strengths encoding weight and distance, as a function of training. Magnitude is given by the range of connection strengths (most positive minus most negative) coming into one weight or distance hidden unit (lower panel) and coming out of weight or distance unit (upper panel).

made to the output weights tend to cancel each other out, and progress of learning is very slow. It is only after the input weights become organized that learning can proceed efficiently on the output side of the hidden units.

The story we are telling would be a very sad one, were it not for the fact that it is not all or none. It is not that there is no learning at all at first. In that case, there would be no gradual change to the point where learning becomes more rapid. Rather, it is simply that initially learning is simply *very* gradual; so gradual that it does not show up in overt behavior. Gradually though this initially slow learning accelerates, producing an increasing readiness to learn.

The differential readiness to learn allows the model to account for the results of an experiment described by Siegler (1976), on the effects of training for young vs. old Rule 1 children. Siegler showed 5- and 8-year-old Rule 1 children a series of distance problems or a series of conflict problems. The children were allowed to try to predict which side would go down, and were then shown what actually happens. The results were striking. Of the children who saw the outcomes for distance problems, both age groups were very likely to exhibit Rule 2 behavior on a post-test. However, of the children who saw the outcomes for conflict training, the younger children either continued to behave in accordance with Rule 1 or became inconsistent in their responding. The older children, on the other hand, benefitted from the conflict training. On a post test, the older children were very likely to exhibit Rule 2 or Rule 3 behavior. In further experiments on early Rule 1 children, Siegler reported that these children do not represent the distance correctly: When asked to reproduce a balance beam configuration, they could usually get the number of weights correct, but could rarely place them on the correct pegs. Younger Rule 1 children who were then trained to represent the distance correctly were able to learn from experience with conflict problems like the older Rule 1 children.

SIMULATIONS OF FEEDBACK EXPERIMENTS

This general pattern of results fits closely with what we would expect based on what we have already learned about the model. Particularly interesting are the effects of feedback on conflict trails on early and late Rule 1 performance. The model was shown a set of conflict trials with feedback, using the weights obtained early in the Rule 1 phase (epoch 20) and later (epoch 40). In the former case, performance gradually reverted to random. In the latter case, it shifts after only one exposure to the set of conflict trials to the rule 2 level. The reason for the deterioration in the first case is simply that early in Rule 1, the weights in the network do not encode distance information at all. As a result the conflict trials appear to involve a pattern of very inconsistent feedback concerning the correct predictions to make based on the weights alone. Later in Rule 1, the distance cue is weakly encoded. While it is still too weak to actually cause the output to be strongly enough affected by the distance cue to actually affect performance, it is strong enough for a small amount of experience to cause a further increase in strength to the point where the distance cue is strong enough to influence performance.

Feedback Simulation Trials

The general approach taken here can be extended to cover the full range of training experiments carried out by Siegler (1976). Simulations more closely matching the design of these experiments were carried out by the second author. For these simulations, a similar architecture was used. One major difference was the addition of 2 more output units. These additional output units received input from the hidden units on the distance dimension. In general, these additional output units, from now on to be referred to as distance encoding units, were not involved in the simulations except where outlined below.

The purpose of these simulations was to model the training experiences from Siegler's second and third experiments. The issue at this point was whether the model could simulate the effects of training with distance problems versus training with conflict problems.

To simulate the stability of the system's knowledge about the weight dimension, the learning mechanism's proportion of change with respect to error for the connections from the weight units and the weight internal units and from the weight internal units to the balance scale output units (their learning rates) were, respectively, .00 and .025, while the proportion of change with respect to error for all of the distance dimension connections was .05.

To begin the simulation of Siegler's second experiment, the model was initialized by training only the connections between the weight input units and the weight-hidden units, and between the weight-hidden units and the balance scale output units. This training was performed by presenting the model with 50 epochs of input and feedback for each possible configuration of weight and balance problems with no distance information provided to the model. On the balance scale prediction task, this system produced perfect Rule 1 behavior. From this base performance level, the model received two types of training in separate sessions. In one session, the model received feedback training similar to the distance feedback training in Siegler's second experiment. That is, the model was trained with 16 different patterns and their associated correct responses; 12 patterns with equal weight input to the two sides of the weight input units and different distance inputs to the two sides of the distance input units (Distance problems), 2 patterns with equal weight and distance inputs (Balance problems), and 2 patterns with different weight inputs to the two sides of the weight input units and equal distance input to the two sides of the distance input units (Weight problems).

In the second session, after reinitialization of the system to the base performance level, the model received feedback training similar to the conflict training in Siegler's second experiment. That is, the model was trained with 16 different patterns and their associated correct responses; 12 patterns with greater weight input to one side of the weight input units and greater distance input to one side of the distance input units (Conflict problems), and 2 patterns with equal weight and distance inputs (Balance problems), and 2 patterns with different weight inputs to the two sides of the weight input units and equal distance input to the two sides of the distance input units (Weight problems).

In both of these sessions, the model received 40 epochs (640 trials) of training. A measure of the effectiveness of the training trials was plotted over the course of the forty training epochs. The effectiveness measure, called the sum of squares error term (SSERROR), measures the difference between the activations of the prediction responses of the model over the training set of patterns and the activations of correct responses for the set of patterns. Therefore, a large error term represents an inability of the model to produce correct predictions, while a small term represents a close match between the predictions of the model and the correct responses.

Figures 3.10a and 3.10b show SSERROR plotted over training epochs during distance training and conflict training, respectively. In Fig. 3.10a, we can see that distance training causes the model's predictive performance over the training problem set to improve dramatically. In contrast, Fig. 3.10b shows that conflict training did not increase the model's predictive performance over the training problem set. These graphs indicate that, in the 40 epoch time frame, the model learned from the distance training but did not learn from the conflict training.

A more dramatic demonstration of this difference in learning was exhibited in the model's performance on the prediction task following the feedback training. Like many of Siegler's five year old subjects, the model learned to perform as a rule 2 user after distance training; in addition to getting balance, weight and conflict-weight problems correct, the model was able to correctly predict distance problems as well. After conflict training, however, the model did not learn to perform at all different from rule 1 behavior. Hence, the model closely simulated the behavior of the 5-year-olds.

The next step in the evaluation of the model was to give distance encoding training to the base performance model. The model's encoding training was not the same as the training that Siegler presented to his subjects; instead, this training was only meant to get the model to categorize the distance input into one of the three relative values. Thus, this training involved modification of only the connections between the distance input units, the hidden units of the distance dimension, and the distance encoding units at the output level. This encoding training set included every configuration of distance as input to the network (36 patterns) paired with the corresponding relative distance value (more distance left, more distance right, or equal distance) as the target feedback for the distance encoding units.

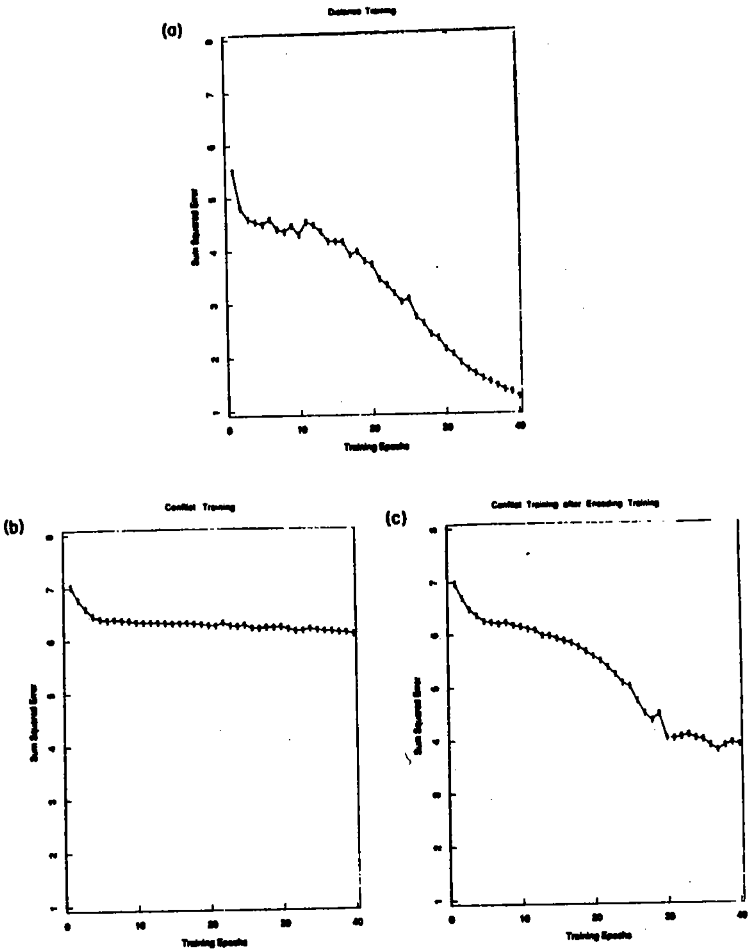


FIG. 3.10. Network error over training for (a) distance training, (b) conflict training, and (c) conflict training after encoding training.

For example, the model was presented at the input level with the activation pattern for three right and two left, and trained to produce the pattern of activation for more distance right across the distance encoding units. The model received 25 epochs of training with this training corpus. This training allowed the distance hidden units to discriminate distance input patterns as greater distance left, greater distance right or equal distance. No training occurred between any of the other units of the model during the distance encoding training phase. Following this encoding training, the model was again tested on the prediction task. The model, like Siegler's subjects, did not exhibit a change in behavior. Finally, the model was provided with the same conflict training as had been provided in the earlier conflict training session. In contrast to Fig. 3.10b, the plot of the effectiveness of conflict training after encoding training (Fig. 3.10c) shows that conflict training dramatically increases the model's predictive performance over the training problem set. This graph indicates that the model was able to learn from the conflict training after encoding training. Following this training, the model was again tested on the prediction task.

This time, the model was able to correctly use the distance information to predict distance problems. It also learned, like the children, not to rely strictly on the weight cues to predict the conflict problems and thus its performance went down on the conflict weight problems (from 100% correct to 75% correct) and up on the conflict distance and conflict balance problems (both from 0% correct to 50% correct).

In essence, the model learned to perform as a rule 3 user after the conflict training following the encoding training. Hence, the model simulated the learning abilities of many of the 5- and 8-year-olds on the conflict training task after receiving distance encoding training.

To understand how this model works, it is important to understand the internal structure that develops during training. So, before distance training, after distance training, after conflict training, and after encoding training followed by conflict training, the model was presented with 15 different distance configurations (patterns of activation) across the distance input units. There were five configurations each; of more distance right, more distance left, and equal distance (balance). After each presentation of a distance configuration, the average activation of the two sets of units of the distance internal group were measured and plotted against each other. The two sets of units were chosen by examining the activation values and attempting to determine which, if any, of the units tended to have correlated activation values. Units with correlated activation values were considered members of a common set for the purposes of this analysis. Units could not be chosen *a priori* due to fact that the learning mechanism recruits units to code similar functions only during the course of training.

To display these activations in response to different inputs, two-dimensional graphs were constructed with each axis specifying the average activation level of one set of units. For each graph, five points are plotted which mark the activations of the two sets of internal units when the five configurations of more distance right were presented—these are marked with an 'R'. Likewise, each of the five points for the internal activations for the configurations of more distance left and balanced distance are marked with an 'L' and a 'B', respectively. Upon examining these activation values before training, it was discovered that these sets of input patterns were not discriminated in any way in the distance internal group. Because the connection strengths of the connections to these units were not trained, every input pattern produced almost the same activation pattern at the internal level and thus all of the points ended up on top of each other. Thus, this graph is not shown. However, when this same procedure was performed after distance training, the activations shown in Fig. 3.11a were produced. Here, we see that the network has learned to internally represent similar distance input similarly (all of the R's, L's, and B's are clustered together), and, more importantly, to represent different distance inputs differently (the R, L, and B clusters are separated away from each other). What is claimed, then, is that during training, the model restructured the internal representations for each problem association in such a way that the concept of relative distance—more distance left, right, and equal distance—emerged from the representations.

This analysis was also performed after conflict training and the results are shown in Fig. 3.11b. In this case, the training did not allow for the formation of the relative distance concept within 40 epochs of training. The internal representations for the different distance configurations did not differentiate into the three

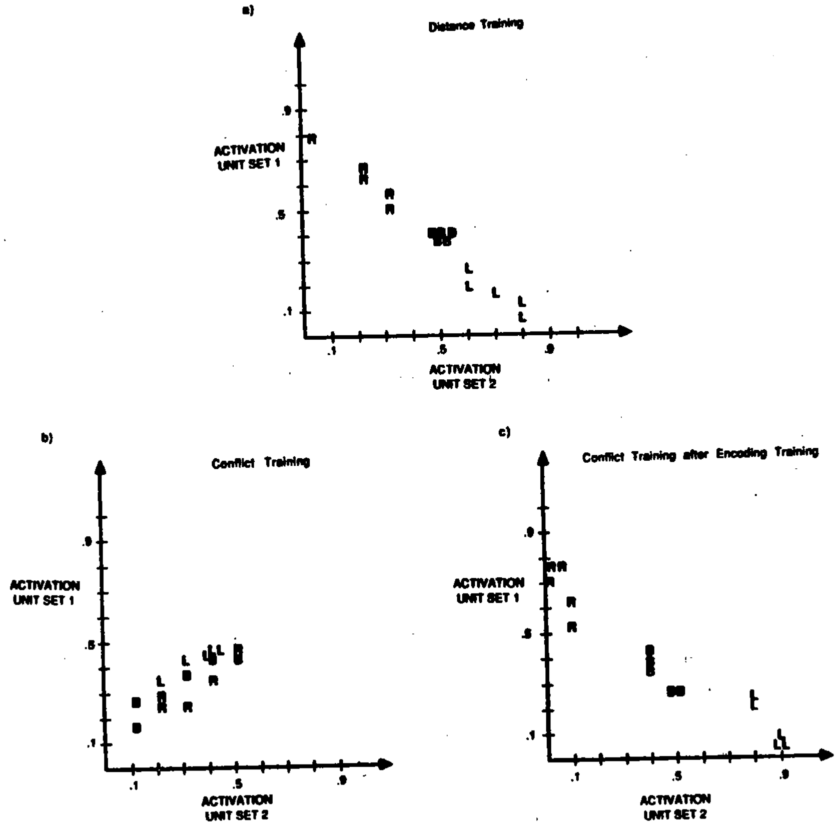


FIG. 3.11. Activations of the units of the internal distance group for (a) distance training, (b) conflict training, and (c) conflict training after encoding training.

clusters of their distance relative relationships. Without this abstract relationship, the model was unable to generalize to new configurations and thus, in the prediction task, did not change its behavior. However, when this analysis was performed after encoding training, (which produced the relative distance representations at the internal level as shown in Fig. 3.11c), the model was able to learn to use relative distance cues on the prediction task.

Shortcomings of the Model

Taken together, the two versions of the simple network model of the balance beam task that we have described above exhibit a striking correspondence with many aspects of the developmental facts. These correspondences have, we think, important implications for theories of cognitive development. Before we turn to these correspondences, we first consider a few shortcomings of the model. Three failures of the first version to fit aspects of Siegler's data must be acknowledged: First, the model can never actually master Rule 4, though some subjects clearly do. Second, it's behavior during Rule 3 is slightly different from humans (though it should be noted that the "human" Rule 3 pattern is actually a mixture of different strategies according to Klahr and Siegler, 1978). Third, it can exhibit position biases that are uncharacteristic of humans, who seem (at least, from the age of 5 on) to "know" that there is no reason to prefer left over right.

There are other shortcomings at well. Perhaps the most serious is in the input representations, which use distinct units to represent different amounts of weight and distance. This representation was chosen because it does not inherently encode the structure of each dimension, thereby forcing the network to discover the ordering of each dimension. But it has the drawback that it prevents the network from extrapolating or even interpolating beyond the range of the discrete values that it has experienced.

Finally, Siegler has reported protocol data that indicates that subjects are often able to describe what they are doing verbally in ways that correspond fairly well to their actual performance. It is not true that all subject's verbalizations correctly characterize the Rule they are using, but it is true, for example, that subjects who are sensitive to the distance cue mention that they are using this cue and those who are not tend not to mention it. The model is of course completely mute.

What are we to make of these shortcomings in light of the overall success of the model? Obviously, we cannot take it as the final word on development of ability to perform the balance scale task. We would suggest that the model's shortcomings may lie in two places: First, in details of the encoding of inputs and of the network architecture; and second, in the fact that the model only deals with acquisition of implicit knowledge.

Regarding the first point, it would be reasonable to allow the input to encode similarity on each dimension by using input representations in which each unit responded to a range of similar values so that neighboring weights and distances produced overlapping input representations; furthermore, the inputs could well make use of a relative code of magnitude to keep values within a fixed range. This would probably overcome the interpolation and extrapolation problems (we have no stand on whether such codings are learned or pre-wired).

These kinds of fixes would not allow the model to truly master Rule 4 and perhaps rightly so, since it seems likely that Rule 4 (unlike the other rules) can only be adhered to strictly as an explicit (arithmetic) rule. Indeed, it must be acknowledged that there is a conscious, verbally accessible component to the problem solving activity that children and adults engage in when they confront a problem like the balance scale problem. The model does not address this activity itself. However, it is tempting to suggest that the model captures the gradual acquisition mechanisms which establish the possible contents of these conscious processes. One can view the model as making available representations of differing salience as a function of experience; these representations might serve as the raw material used by the more explicit reasoning processes that appear to play a role. This is of course sheer speculation at this point. It will be an important part of the business of the ongoing connectionist exploration of cognitive development to make these speculations explicit and testable.

IMPLICATIONS OF THE BALANCE SIMULATIONS

The model captures several of the more intriguing aspects of cognitive development. It captures its stage-like character, while at the same time exhibiting an underlying continuity which accounts for gradual change in readiness to move on to the next stage. It captures that fact that behavior can often seem very much to be under the control of very simple and narrow rules (e.g., Rule 1), yet exhibit symptoms of gradedness and continuity when tested in different ways. It captures the fact that development, in a large number of different domains, progresses from an initial over-focussing on the most salient dimension of a task or problem—to the point where other dimensions are not even encoded—followed by a sequence of further steps in which the reliance on the initially unattended dimension gradually increases.

As mentioned previously, the model can be seen as implementing the accommodation process that lies at the heart of Piaget's theory of developmental change. Accommodation essentially amounts to adjusting mental structures to reduce the discrepancy between observed events and expectations derived from the existing mental structures. According to Flavell (1963), Piaget stressed the continuity of the accommodation process, in spite of the overtly stage-like character of development, though he never gave a particularly clear account of how stages arise from continuous learning (see Flavell, 1963, pp. 244–249 for a description of one attempt). The model provides such a description: it shows clearly how a continuous accommodation-like process can lead to a stage-like progression in development.

Changes in representation and attention through the course of development. When a balance scale problem is presented to the model, it sees it in different ways, depending on its developmental state. At all times, information is in some sense present in the input for determining what is the correct response. However, at first this information produces no real impression; weak, random activations occur at the hidden level and these make weak, random impressions at the output level. At the beginning of the Rule 1 behavioral phase, the model has learned to represent relative amount of weight. The pattern of activation over the hidden units captures relative weight, since one unit will be more activated if there is more weight to the right, and the other will be more activated if there is more weight to the left; both units take on intermediate activations when the weights balance. At this point, we can see the model as encoding weight, but not distance information. Indeed, as we have seen at this point the network could be said to be ignoring the distance cue; it makes little impact on activation, and learning about distance is very slow at this point. At the end of the Rule 1 phase, in spite of its lack of impact on overt behavior, the network has learned to represent relative distances; at this point it is extremely sensitive to feedback about distance; it is ready to slip over the fairly sharp boundary in performance between Rule 1 and Rule 2. Thus, we can see the Rule 1 stage as one in which overt behavior fails to mirror a gradual developmental progression that carries the model from extreme unreadiness to learn about distance at the beginning of this phase to a high degree of readiness at the end.

This developmental progression seems to resolve the apparent paradoxical relation between observed stage-like behavioral development and assumed continuity of learning. To us this is the most impressive achievement of the model; it provides a simple, explicit alternative to maturational accounts of stage-like progression in development.

It must be noted, however, that the success of the model depends crucially on its structure. In fact the results are less compelling if either of the following changes are made: (a) if balance is treated as a separate category, rather than being treated as the intermediate case between left-side-down and right-side-down; (b) if the connections from input to hidden units are not restricted as they are here so that weight is processed separately from distance before the two are combined.

More generally, it is becoming clear that architectural restrictions on connectionist networks are crucial if they are to discover the regularities we humans discover from a limited range of experiences (Denker et al., 1987; Rumelhart, in preparation). This observation underscores that fact that the learning principle, in itself, is not the only principle that needs to be taken into account. There probably are additional principles that are exploited by the brain to facilitate learning and generalization. Just what these additional principles are and the extent to which they are domain specific remains to be understood in more detail.

Extending this observation a step further, we can see the connectionist framework as a new paradigm in which to explore basic questions about the relations of nature and nurture. We may find that successful simulation of developmental processes depends on building in domain specific constraints in considerable detail; if so this would support a more nativist view of the basis of domain–specific skills. On the other hand, it may turn out that a few other general principles in addition to the learning principle are sufficient to allow us to capture a wide range of developmental phenomena. In this case we would be led toward a much more experience-based description of development. In either case, it seems very likely that connectionist models will help us take a new look at these important basic questions.

CONCLUSIONS

The exploration of connectionist models of human cognition and development is still at an early stage. Yet already these models have begun to capture a new way of thinking about processing, about learning and, we hope the present chapter shows, about development. Several further challenges lie ahead. One of these is to build stronger bridges between what might be called cognitive–level models and our evolving understanding of the details of neuronal computation. Another will be to develop more fully the application of cognitive models to higher-level aspects of cognition. The hope is that the attempt to meet these and other challenges will continue to lead to new discoveries about the mechanisms of human thought and the principles that govern their operation and adaptation to experience.

ACKNOWLEDGMENTS

The text of this article overlaps substantially with the text of an earlier article by McClelland (1989) and contains excerpts from an article by Jenkins (1986). The work reported here was supported by ONR and DARPA Under Contract Number N00014-86-K-0349, by an NIMH Career Development Award MH00385 to the first author, by ONR Contract N00014-82-C-0374, and by NSF Grant BNS 88-12048.

REFERENCES

Bever, T. G. (1970). The cognitive basis for linguistic structures. In J. R. Hayes (Ed.), *Cognition and the development of language* (pp. 279–362). New York: Wiley.

Denker, J., Schwartz, D., Wittner, B., Solla, S., Hopfield, J., Howard, R., & Jackel,

L. (1987). *Automatic learning, rule extraction, and generalization* (AT&T Bell Labs Technical Report). Holmdel, NJ: AT&T Bell Labs.

Feldman, J. A. (1988). Connectionist representation of concepts. In D. Waltz & J. A. Feldman (Eds.), *Connectionist models and their implications: Readings from cognitive science* (pp. 341–363). Norwood, NJ: Ablex.

Ferretti, R. P., Butterfield, E. C., Cahn, A., & Kerkman, D. (1985). The classification of children's knowledge: Development on the balance scale and inclined plane tasks. *Journal of Experimental Child Psychology, 39,* 131–160.

Flavell, J. H. (1963). *The developmental psychology of Jean Piaget.* Princeton, NJ: D. Van Nostrand.

Grossberg, S. (1978). A theory of visual coding, memory, and development. In E. L. J. Leeuwenberg & H. F. J. M. Buffart (Eds.), *Formal theories of visual perception.* New York: Wiley.

Hinton, G. E. (1989). Learning distributed representations of concepts. In R. G. M. Morris (Ed.), *Parallel distributed processing: Implications for psychology and neurobiology* (pp. 46–61). New York: Oxford University Press.

Hinton, G. E. (in press). Connectionist learning procedures. *Artificial Intelligence.*

Hinton, G. E., McClelland, J. L., & Rumelhart, D. E. (1986). Distributed Representations. In D. E. Rumelhart, J. L. McClelland, & the PDP research group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition. Volume I.* Cambridge, MA: Bradford Books.

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

Jenkins, E. A., Jr. (1986). Readiness and learning: A parallel distributed processing model of child performance. Pittsburgh, PA: Carnegie-Mellon University, Psychology Department.

Keil, F. C. (1979). *Semantic and conceptual development: An ontological perspective.* Cambridge, MA: Harvard University Press.

Klahr, D., & Siegler, R. S. (1978). The representation of children's knowledge. In H. W. Reese & L. P. Lipsitt (Eds.), *Advances in child development and behavior* (pp. 61–116). New York: Academic Press.

McClelland, J. L. (1985). Putting knowledge in its place: A scheme for programming parallel processing structures on the fly. *Cognitive Science, 9,* 113–146.

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

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: Part 1. An account of basic findings. *Psychological Review, 88,* 375–407.

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

McClelland, J. L., Rumelhart, D. E., and the PDP research group. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition. Volume II.* Cambridge, MA: Bradford Books.

McClelland, J. L., St. John. M., & Taraban, R. (1989). Sentence comprehension: A parallel distributed processing approach. *Language and Cognitive Processes, 4,* 287–335.

Patterson, K., Seidenberg, M. S., & McClelland, J. L. (1989). Connections and disconnections: Acquired dyslexia in a computational model of reading processes. In R. G. M. Morris (Ed.), *Parallel distributed processing: Implications for psychology and neurobiology* (pp. 131–181). New York: Oxford University Press.

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

Rumelhart, D. E. (in preparation). *Generalization and the learning of minimal networks by back propagation.*

Rumelhart, D. E. (1990). Brain style computation: Learning and generalization. In S. Zornetzer, J. Davis, & C. Lau (Eds.), *An introduction to neural and electronic networks* (pp. 405–420). New York: Academic Press.

Rumelhart, D. E., & McClelland, J. L. (1986). On learning the past tenses of English verbs. In J. L. McClelland, D. E. Rumelhart, & the PDP research group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition. Volume II.* Cambridge, MA: Bradford Books.

Rumelhart, D. E., & Norman, D. A. (1982). Simulating a skilled typist: A study of skilled cognitive-motor performance. *Cognitive Science, 6,* 1–36.

Rumelhart, D. E., Hinton, G. E., & McClelland, J. L. (1986). A framework for parallel distributed processing. In D. E. Rumelhart, J. L. McClelland, & the PDP research group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition. Volume I.* Cambridge, MA: Bradford Books.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart, J. L. McClelland, & the PDP research group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition. Volume I.* Cambridge, MA: Bradford Books.

Rumelhart, D. E., McClelland, J. L., & the PDP research group. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition. Volume I.* Cambridge, MA: Bradford Books.

St. John, M. F., & McClelland, J. L. (1988). *Learning and applying contextual constraints in sentence comprehension.* (AIP Technical Report). Pittsburgh, PA: Carnegie Mellon University, Departments of Computer Science and Psychology, and University of Pittsburgh, Learning Research and Development Center.

Sejnowski, T. J., & Rosenberg, C. R. (1987). Parallel networks that learn to pronounce English text. *Complex Systems, 1,* 145–168.

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

Siegler, R. S. (1981). Developmental sequences within and between concepts. *Monographs of the Society for Research in Child Development, 46* (No. 189, pp. 1–74).

Siegler, R. S., & Klahr, D. (1982). When do children learn? The relationship between existing knowledge and the acquisition of new knowledge. In R. Glaser (Ed.), *Advances in instructional psychology, Vol. 2* (pp. 121–211). Hillsdale, NJ: Lawrence Erlbaum Associates.

Wilkening, F., & Anderson, N. H. (in press). Representation and diagnosis of knowledge structures. In N. H. Anderson (Ed.), *Contributions to information integration theory.*

Williams, R. J. (1986). The logic of activation functions. In D. E. Rumelhart, J. L. McClelland, & the PDP research group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition. Volume I.* Cambridge, MA: Bradford Books.