4

A Connectionist Perspective on Knowledge and Development

J. L. McClelland
Carnegie Mellon University

Questions about how our knowledge changes in response to experience lie at the heart of efforts to understand cognitive development. In this chapter, I approach these questions from a connectionist perspective. I contrast a connectionist approach to these questions with traditional symbolic or propositional approaches. I suggest that thinking about the development of knowledge has been heavily influenced by the assumption that knowledge is symbolic, and I argue that a connectionist approach leads to new conceptualizations of the processes through which developing children come to know more and more about the world.

These issues are explored by considering a connectionist simulation model that is applied to the balance scale task studied by Siegler and others. The graded nature of the representations used by the model allows it to account for several aspects of the empirical data, including the Torque Difference Effect (Ferretti & Butterfield, 1986).

The incremental nature of connectionist learning—the fact that current learning builds on what has already been learned—allows the model to account for stagelike developmental progressions and for differences in readiness to learn from particular experiences at different points in development. The chapter also shows how the connectionist framework allows one to capture effects of cue complexity as well as cue familiarity on the course of development. The discussion considers the essential features of the connectionist account of performance and development in the balance scale task, and considers open questions, such as the nature of the initial constraints necessary to lead to successful development, and the relation-

ship between the implicit knowledge that is captured by connectionist models and explicit knowledge such as verbalizable propositions and rules.

WHAT IS KNOWLEDGE?

Let us begin with the fundamental question: What is knowledge, anyway? According to the symbolic approach, knowledge takes two forms: a set of propositions, involving specified relations among specified symbols standing for objects or classes of objects; and a system of rules for using knowledge to make inferences and guide actions. Propositional knowledge can be acquired through encoding experienced events into propositional form and through inferences applied to encoded propositions. Thus, if I know *All men are mortal*, and I learn through experience or direct instruction that *Socrates is a man*, and if I know the appropriate rule of inference, then I can infer that *Socrates is mortal*. Much theoretical work in developmental psychology explicitly or implicitly adopts this symbolic approach without questioning it. Thus Siegler, in his seminal papers on development in the balance scale task, characterized children's knowledge in terms of a set of rules; Spelke, Breinlinger, McComber, and Jacobson (1992) characterized infants' knowledge of intuitive physics in terms of innate principles with which children reason; and Pinker (1991) characterized children's knowledge of morphology in terms of a simple rule system, complemented by a separate associative system used for exceptions.

This chapter explores the view that much of the knowledge that developmental psychologists study may not be propositional. Instead, I suggest the knowledge may be stored in the form of connections: that is, graded parameters embedded in specific processing structures that use them. This conception of the nature of knowledge itself leads to a change in thinking about how knowledge is acquired; not by inference as in the symbolic case, but by gradual parameter adjustment. I do not mean to suggest that no knowledge is symbolic or that no discovery of new knowledge occurs by inference; I only mean to argue that the knowledge that underlies children's performance in many developmental tasks may have this graded, embedded, nonsymbolic character.

To begin our exploration of this connectionist approach, it is useful to start with an overview of the connectionist framework. The framework is now quite familiar (see Rumelhart, McClelland, and the PDP Research Group, 1986, for an introduction), so the overview is brief. On this approach—also sometimes called the parallel-distributed processing (PDP) approach—information processing takes place through the interactions of large numbers of simple, neuronlike processing units, arranged into modules. An active representation—such as the representation one may have of a current perceptual situation, for example, or of an appropriate overt

response—is a distributed pattern of activation, over several modules, representing different aspects of the event or experience, perhaps at many levels of description. Processing in such systems occurs through the propagation of activation among the units, through weighted excitatory and inhibitory connections.

As already noted, the knowledge in a connectionist system is stored in the connection weights: it is the connections that determine what representations we form when we perceive the world and what responses these represenations will lead us to execute. Such knowledge has several essential characteristics: (a) it is incohate, implicit, and completely opaque to verbal description; (b) even in its implicit form it is not necessarily accessible to all tasks—rather, it can be used only when the units it connects are actively involved in performing the task; (c) it can arbitrarily approximate symbolic knowledge but it need not—it admits of states that are cumbersome at best to describe by rules; and (d) its acquisition can proceed gradually, through a simple, experience-driven process. At certain times during acquisition, knowledge may be approximately characterizable in terms of one or another system of symbolic rules, but transition between such states of knowledge may be completely seamless, governed by a completely homogeneous learning process.

Let us consider the learning process in more detail, because it is the heart of the process of developmental change in connectionist systems (McClelland, 1989). Various approaches to learning have been taken within the PDP framework, but the one that appears to be most promising for understanding cognitive development is a procedure that learns from the mismatch between expected and observed events. In this approach, we imagine that the cognitive system is continually engaged in making implicit predictions for the immediate future, based on its represention of the current situation (cf. Rescorla & Wagner, 1972). The representation of the current situation is a pattern of activation over a set of internal units, and the prediction is represented as a pattern of activation over a set of output units. These predictions are compared to a pattern that represents what actually happens in the world, and the dicrepancy is used to adjust the weights. The actual rule for connection strength adjustment takes the following form:

> *Adjust each parameter in proportion to the extent that its adjustment will reduce the discrepancy between predicted and observed events.*

This is equivalent to a procedure for adjusting connection weights:

> *Adjust each connection weight in proportion to the extent that its adjustment will reduce the discrepancy between the output the network produces and the desired output specified by the environment.*

This approach to learning in connectionist systems was pioneered by Rosenblatt (1959) and Widrow and Hoff (1960); the generalization, known as *backpropagation*, was developed by Rumelhart, Hinton, and Williams (1986). These procedures perform a search process called *gradient descent:* The process of connection adjustment is seen as a process of search across a surface in a large space, in which the height of the surface represents the error, and in which the surface is defined over a large number of other dimensions, one for each connection weight. Each point in the space represents a possible entire set of connection weights and the corresponding error, and from each point there is one direction that represents the steepest direction downhill in the error measure. This direction is called the *gradient* (it represents the negative of the slope of the error surface at that point), and gradient descent simply amounts to moving down the gradient. It is useful to define the gradient in terms of an entire ensemble of possible events and experiences in the environment. In this case, each particular event gives a random sample of the gradient, rather than a true picture of the entire gradient. If we adjust connection weights based on this sample, the learning procedure is more properly called *stochastic gradient* to indicate that learning is based, not on the exact gradient, but on a random sample of it (see White, in press, for a discussion).

In this chapter I explore the effects of using the stochastic gradient approach to learning in connectionist systems that are exposed to environments that exhibit regularities in the predictions that can be made from representations of certain situations to subsequent outcomes. I show how this approach offers a new way of thinking, not only about the knowledge that underlies performance in cognitive tasks, but also about the process of developmental change. And I demonstrate that the approach has considerable appeal in accounting for a wide range of findings obtained in studies based on the balance scale task used by Siegler (1976, 1981) and others. I show how the connectionist approach is consistent with a considerable body of recent evidence on the graded nature of the knowledge children use in making cognitive judgments in this task. I also show that the connectionist approach can lead us to understand why there are periods of relative stasis in development, punctuated by periods of relatively rapid change. I discuss how the approach can lead us to understand how readiness to profit from particular experience may change gradually as a child performs overtly at the same developmental level over an extended period of time. The choice of the balance scale task allows us to compare the connectionist approach to the symbolic approach taken in work by Siegler (Klahr & Siegler, 1978; Siegler, 1976, 1981; Siegler & Klahr, 1982) and to the algebraic approach taken by Wilkening and Anderson (1991). Some of the connectionist simulation work reviewed here was reported in McClelland (1989) and

McClelland and Jenkins (1991). However, I extend the previous simulations to address the torque difference effect of Ferretti and Butterfield (1986) and to examine factors that influence ease of mastery of the weight and distance cues that must be used to perform correctly in the balance scale task.

THE BALANCE SCALE TASK

The balance scale task was introduced by Inhelder and Piaget (1958) and studied extensively by Siegler (1976, 1981) and many others. In the standard version of the task, which is the main focus here, the child is presented with a balance scale like the one in Fig. 4.1. Some number of weights are placed on one peg on the left of the fulcrum, and some number of weights are placed on one peg on the right. The child's task is to predict which side would go down if the scale were free to move. Typically a series of trials is given with different numbers of weights on different pegs, and there is no feedback; that is, the scale is imobile so that the child does not learn whether the prediction is right or wrong.

Siegler's Rules

Siegler (1976) developed a set of possible rules that children might use in the balance scale task, and a procedure for determining which of the rules the child was using. The rules, taken from Siegler (1976), are presented in Fig. 4.2. A quick summary can be given as follows: Children who use Rule 1 attend to the number of weights on each side, but not the the distance from the fulcrum. Thus, they say the sides balance if the weights are the same on both sides; otherwise, they say the side with the greater weight will go down. Children who use Rule 2 are like children who use Rule 1, except that they take distance into account if the weights are the same on both sides; in this case they say the side where the weights are the furthest from the fulcrum will go down. Children who use Rule 3 appreciate that both weight and distance matter. For these children, if the number of weights is greater on one side and the distance is greater on the other, the child will be uncertain

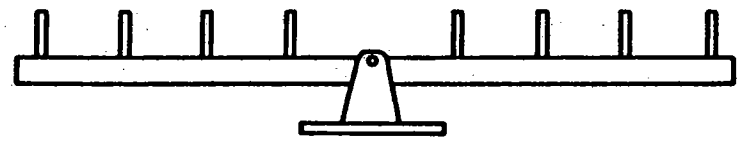

FIG. 4.1. A balance scale of the type used by Siegler (1976, 1981). Reprinted from Figure 1 of Siegler (1976), with permission.

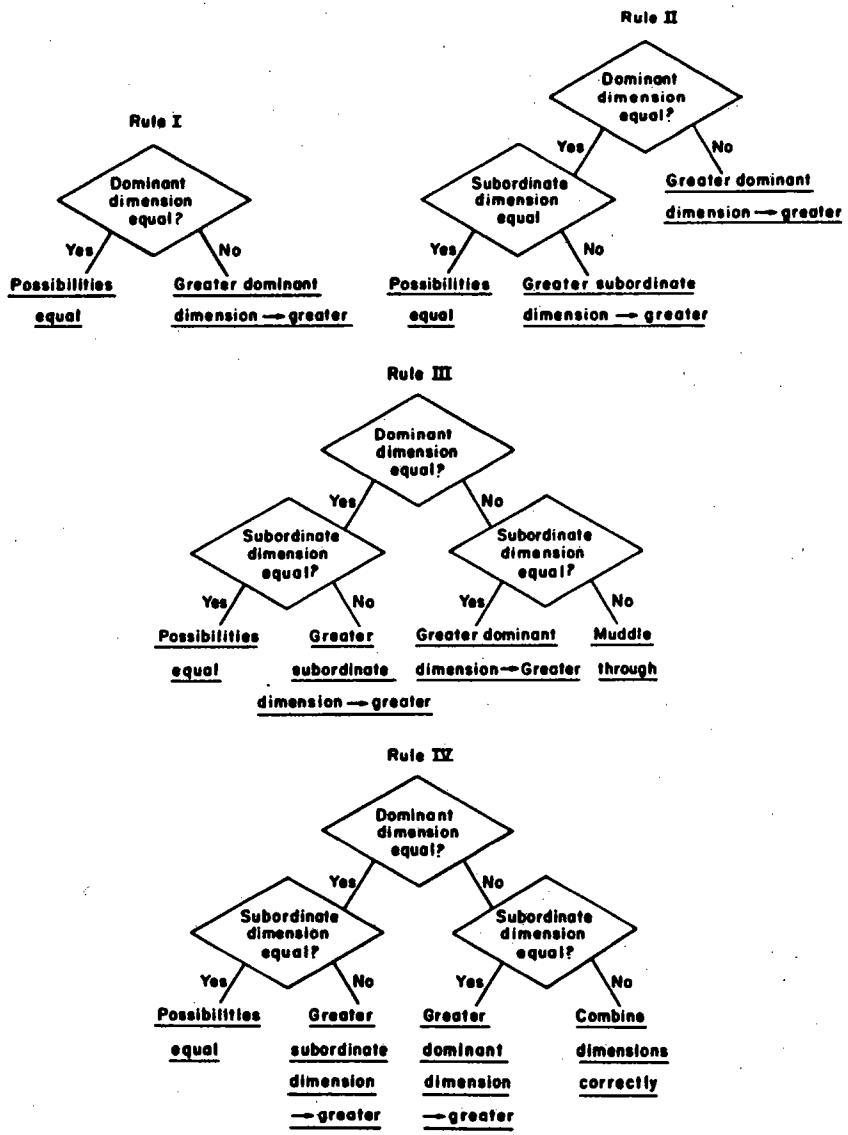

FIG. 4.2. The four Rules identified by Siegler (1976). Reprinted from Figure 1 of Siegler (1981), with permission.

what to do; operationally, the assumption is that the child simply guesses, distributing the guesses evenly between saying (a) the side with the greater weight goes down, (b) the side with the greater distance goes down, or (c) the sides balance. Children who use Rule 4 appreciate that both cues matter, as well; they differ from children who use Rule 3 in that they understand the

physical torque principle that governs which side will go down. This principle implies that the side where the product of weight times distance is greater will go down, and they use this rule in case the cues are in conflict. This allows correct performance in every case.

To assess conformity to these rules, Siegler developed a test consisting of four examples of each of six problem types (Fig. 4.3). In *balance* problems, the weight and the distance were the same on both sides. In *weight* problems, only the weight differed. In *distance* problems, only the distance differed. In the remaining three problem types, both weight and distance differed, and both cues were always in conflict, so that the distance was greater on one side but the weight was greater on the other. For *conflict-weight* problems, the torque was greater on the side with the greater weight; for the *conflict-distance* problems, the torque was greater on the side with the greater distance; and for the *conflict-balance* problems, the torque was the same on both sides. Fig. 4.3 indicates the pattern of responding predicted from each rule for each problem type.

PREDICTIONS FOR PERCENTAGE OF CORRECT ANSWERS AND ERROR PATTERNS ON POSTTEST FOR CHILDREN USING DIFFERENT RULES

Problem type	Rules I	Rules II	Rules III	Rules IV	Predicted developmental trend
Balance	100	100	100	100	No change-allchildren at high level
Weight	100	100	100	100	No change-all children at high level
Distance	0 (Should say "balance")	100	100	100	Dramatic improvement with age
Conflict-weight	100	100	33 (Chance responding)	100	Decline with age Possible upturn in oldest group
Conflict-distance	0 (Should say "right down")	0 (Should say "right down")	33 (Chance responding)	100	Improvement with age
Conflict-balance	0 (Should say "right down")	0 (Should say "right down")	33 (Chance responding)	100	Improvement with age

FIG. 4.3. Examples of each of the six problem types and patterns of performance that would be predicted by each of the six rules. Reprinted from Table 1 of Siegler (1976), with permission.

Siegler's Findings

Over a series of studies, Siegler (1976, 1981) found that the behavior of about 93% of the subjects aged 5 and up conformed to the predictions of one of the four rules. Scoring was fairly strict, but not absolutely so: 20 out of 24 of the subject's responses had to correspond to a rule before the child was said to conform to it, but this meant that up to 4 responses could be deviant. Fig. 4.4 presents the actual profiles of children who were said to conform to each rule, together with the predicted pattern based on the rule. (Also shown are the predictions of the model to be described later.) There is a fairly close correspondence between the rules and children's behavior, but there are discrepancies that may be at least somewhat systematic; I shall have more to say about these when I consider the predictions of the connectionist model. In one study, children were tested twice to assess the reliability of the rule assessment procedure. In general, consistency was high, although it was not perfect; in particular, children scored as using Rule 2 at the first test showed considerable variability at the subsequent test.

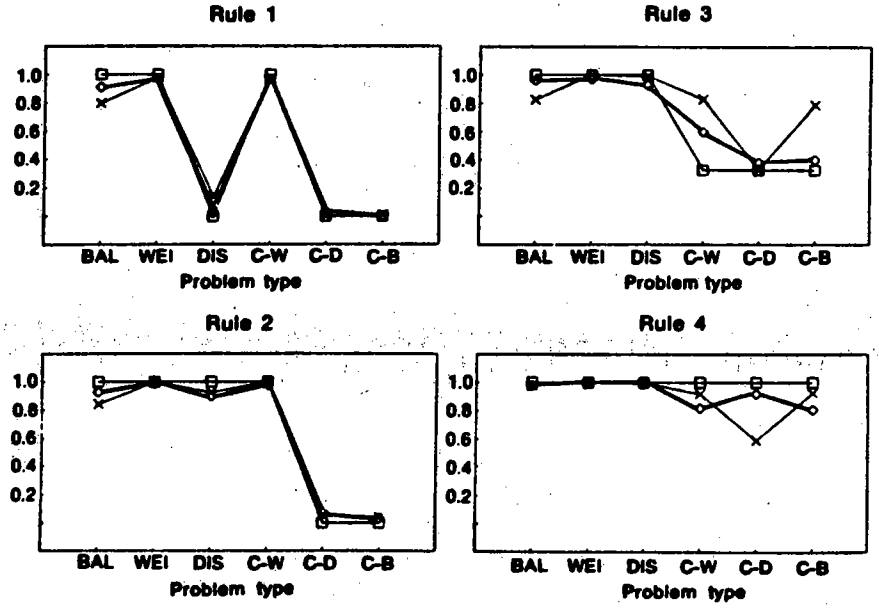

FIG. 4.4. Patterns of performance on each of the six problem types by children (circles), and the McClelland (1989) balance scale model (Xs) averaged over children or networks satisfying Siegler's (1981) criteria for use of each rule. Also shown is the pattern of performance corresponding to exact adherence to each of Siegler's rules (squares). Reprinted from Figure 2.10 of McClelland (1989), with permission.

Considering developmental trends, there was a strong trend for young children (ages 5–7) to conform to Rule 1 and for children beyond the age of 10 or so to conform to Rule 3 or 4. Rule 2 was used by the fewest children, mostly around the age of 7 to 9 years old. Note that not all young adults use Rule 4; even in groups of college undergraduates, the rate of use of Rule 4 was far from perfect. Also, if the results for so-called Rule 4 subjects shown in Fig. 4.4 indicate that these subjects were not completely consistent in their handling of Rule 4.

To summarize these results, there appears at first glance to be a striking conformity between children's behavior and Siegler's rules, and a clear developmental trend to progress from Rule 1 to Rule 3 and sometimes to Rule 4, with Rule 2 serving as a (possibly optional) transitional rule between Rules 1 and 3. Yet, throughout the data, we see some discrepancies. Not all children score in accordance with any rule (and it seems doubtful that their responses were random). Of those who do score in accordance with a rule, it is clear that frequently not all of their responses match.

Underlying Continuity?

Several other researchers have studied the balance scale task or variants of it and have obtained additional results that suggest that the picture painted by the rule assessment method may not capture all aspects of the relevant knowledge possessed by children. There are two main discoveries. First, the rule that best fits a particular child depends on details of the problems used to assess the rules (Ferretti & Butterfield, 1986; Ferretti, Butterfield, Cahn, & Kerkman, 1986). Second, there are alternative rules or procedures that children might use that could yield data that masquarade as one of Siegler's rules (Wilkening & Anderson, 1982, 1991). Wilkening and Anderson (1991), using a functional measurement approach, and Ferretti et al., using the rule assessment method but with a wider range of problems of each type to permit a more detailed examination, suggested that response patterns that show up on Siegler's test as indicative of the use of any of Rules 1 to 4 sometimes reflect the use of an algebraic rule that incorporates graded influences of weight and distance.

Following their lead, one can construct a decision procedure in which one computes a "psychological torque" T (this need not correspond to true physical torque) for each side of the scale:

$$T_s = \chi\Omega_s + \gamma\Lambda_s + \zeta\Omega_s\Lambda_s \quad (1)$$

where subscript s indexes the two sides of the scale (l and r), Ω_s and Λ_s are psychological weight and distance variables, and χ γ, and ζ are parameters

of the child's knowledge. One then chooses a response by computing the difference between T_l and T_r, choosing the side with greater T if the absolute value of the difference is greater than or equal to some criterion C, and choosing *balance* as the response, otherwise. If we make the simplifying assumption that Ω_s is proportional to the actual number of unit weights W_s and Λ_s is proportional to the actual number of units of distance D_s as defined by the experimenter, then the equation can be rewritten:

$$T_s = xW_s + yD_s + zW_sD_s \qquad (2)$$

where x, y, and z are proportional to χ, γ and ζ.

Particular choices of x, y, z, and C now allow us to mimick all of Siegler's rules:

Rule 1. We get exact equivalence to Rule 1 if $y = z = 0$, $C > 0$, *and* $x \geq C$. For example, suppose we choose $x = 1$ and $C = 0.5$. Then Equation 2 reduces to:

$$T_s = W_s \qquad (3)$$

So, if the weights are the same on both sides, the child will say balance (the difference is equal to 0, therefore less than C), but if the number of weights differs, the criterion will be exceeded (minimum difference given unit weights is 1, which is greater than C), and the child will say that the side with the greater weight goes down.

Rule 4. We get equivalence to Rule 4 if $x = y = 0$, $C > 0$, *and* $z \geq C$. For example, $z = 1$ and $C = 0.5$ produces exact conformity to Rule 4.

Rule 2. We can produce conformity to Rule 2 under Equation 2 only if we restrict the range of possible differences in distance. Suppose that the maximum difference in distance between the two sides is M. Then we can implement Rule 2 by choosing $z = 0$, $x > My$, and $y > C$. For example, if the maximum difference in distance is 5, we can choose $x = 6$, $y = 1$, and $C = 0.5$. What this amounts to is the assumption that the weight cue is much stronger than the distance cue, and so, if there is any difference in weight it "outweighs" the largest possible difference in distance. Of course, if children who match Rule 2 were really using Equation 2 with these parameters, then there would be some difference-of-distances that would lead them to choose the side with greater distance, even if there is a slight asymmetry of weight.

Rule 3. Note that Rule 3 as defined by Siegler is meant to encompass any strategy in which both weight and distance influence performance but a strict computation of torque is not used. Given the particular problems used by Siegler (1981), kindly provided to me by Siegler (personal communication, October, 1993), it turns out that the simple rule of choosing the side with the greater sum of weight and distance ($x = y = 1$, $z = 0$, $C = 0.5$) results in a pattern of 2 errors out of the 4 conflict problems of each type. This pattern is categorized as an example of the Rule 3 pattern. Likewise, many other additive or mixed additive and multiplicative compensatory strategies will produce Rule 3 behavior.

Matters become even more complex when we consider the fact that there are broad ranges of the space of possible values of the parameters x, y, z, and C that would produce approximate adherence to one or another of Siegler's rules for a particular set of problems. Points in the parameter space that are outside the regions that allow pure rule emulation often allow an adequate approximation to the rule to be categorized under it, given the leniency of the scoring procedure and the restricted range of examples used in particular cases.

The Torque Difference Effect

The algebraic model's use of graded parameters allows it to address the torque difference findings of Ferretti and Butterfield (1986). These investigators constructed sets of problems of the same six types as those used by Siegler, but they explicitly varied the magnitude of the difference in torque between the two sides of the balance scale. There were four levels, where level 1 corresponded to the most minimal torque difference possible between the two sides of the scale, and level 4 corresponded to the largest difference possible within the confines of the problem space (one to six weights on one peg on each side of a fulcrum, with pegs located from one to six distance units from the fulcrum). Each subject was tested with four weight, distance, conflict-weight, and conflict-distance problems at each level of torque difference, as well as a common set of balance and conflict-balance problems (for these two types of problems, torque difference is fixed at 0). This allowed them to examine both the effect of the torque difference variable on children's performance on problems of particular types, and to score children's adherence to each of Siegler's rules separately for each level of torque difference. There were two principal findings. First, the probability of responding correctly was strongly influenced by torque difference, particularly for distance and conflict-distance problems. In both cases, the probability of correct responding increased substantially as torque difference increased. There were slight effects on

probability of correct responses for weight and conflict-weight problems, but performance on problems of these types was quite good (85% correct) even at the lowest level of torque difference, and there was a more limited range available for improvement.

The second finding was that apparent adherence to Siegler's rules differed at different levels of torque difference. The data are shown in Table 4.1. As torque difference increased, the percentage of children classified as using Rule 1 decreased, and the percentage classified as Rule 4 increased. The percentage of children classified as using Rule 2 increased and then decreased, and there was a similar, but weaker trend for Rule 3.

These results must be interpreted cautiously, because at the larger torque differences used in this study, a variety of different strategies would allow correct responding on conflict-weight and conflict-difference problems. This means, for example, that the subject may be able to get all of the large torque-difference problems correct without actually multiplying weight times distance and comparing torques, as Seigler's Rule 4 requires (in Siegler, 1976, 1981, care was taken in constructing the conflict problems to prevent apparent success for children using some possible nonmultiplicative strategies). Other aspects of the data are not susceptible to this particular problem, however. If a child were really using Rule 1 or Rule 2 as stated by Siegler, that child's classification would not be affected by torque difference, and yet there were substantial effects of that variable on the probability that children were classified as using either of these two rules.

The torque difference effect is consistent with the idea that the underlying procedures used by children may make use of graded information, in accordance with the algebraic model previously given. The algebraic model, however, has some limitations. It can describe a child's developmental state in terms of the values of a few parameters, but it provides no mechanism for change. What is needed is a model that not only captures the developmental state of a child, but at the same time allows us to account for the process of change of state. We now consider one important and interesting aspect of this process: differential readiness to profit from experience at different points in development.

TABLE 4.1
Percentage of Rule Classifications at Different Torque-Difference (TD) Levels (Ferretti & Butterfield, 1986)

TD Level	*1*	*2*	*3*	*4*
1	.29	.19	.17	.05
2	.24	.34	.14	.08
3	.22	.31	.22	.10
4	.19	.15	.15	.37

Readiness

Differential readiness was exhibited in Siegler's work on the balance scale in a series of studies contrasting 5- and 8-year-olds who both scored as Rule 1 users (Siegler, 1976).

In the study of greatest interest here, groups of 5- and 8-year-old Rule 1 users were given a series of 16 conflict problems, with feedback. The children were shown the problem, with the two sides of the scale immobilized. They were then asked to predict which side would go down, and after their prediction the scale was freed so that they could see the actual outcome. The results were quite different for the two groups: Most of the 8-year-olds advanced from use of Rule 1 to a more sophisticated rule (Rule 2 or 3). However, none of the 5-year-olds advanced; half continued to perform at the Rule 1 level, and the other half became unclassifiable, failing to conform to any of the rules (Table 4.2).

Follow up experiments by Siegler (1976) suggested a difference between 5- and 8-year-old children that could account for the difference between the two groups. He asked 5- and 8-year-old children to reproduce balance scale configurations provided by an experimenter. Although 8-year-olds reproduced weight and distance from the fulcrum equally well, 5-year-olds failed to reproduce the distance cue. Through several studies, Siegler established that 5-year-olds cannot encode distance when explicitly instructed to do so; for them to encode distance reliably, they must be given explicit instruction in how to encode it. This strongly suggests that one of the developmental differences between 5- and 8-year-olds is that the 5-year-olds lack not just the inclination, but the ability to encode distance from the fulcrum.

Within the context of the symbolic rule approach, Siegler's results suggest that 8-year-olds do spontaneously encode distance; but those 8-year-olds

TABLE 4.2
Conformity to Siegler's Rules by 5- and 8-year-old Subjects Initially Conforming to Rule 1 After Exposure to Conflict Problems

1. Children not given explicit training in encoding distance				
Age	*1*	*2*	*3*	*Unclass.*
5	5	0	0	5
8	0	2	5	3
2. Children who were given pretraining in encoding distance				
Age	*1*	*2*	*3*	*Unclass.*
5	1	3	4	2
8	0	3	7	0

Note. All subjects were scored as conforming to Siegler's Rule I before training.

who adopt Rule 1 in the balance scale task do not spontaneously use this cue in making judgments. However, they can be induced to use it if given feedback indicating that predictions made simply from the weight cue are incorrect. A further study demonstrated that if 5-year-olds are explicitly instructed in how to encode distance, they can do so. Furthermore, the training was sufficient to allow these children to then profit from exposure to a series of conflict problems.

These results suggest that the tendency to spontaneously encode the distance cue accounts for the difference between early (5-year-old) and late (8-year-old) Rule 1 children. But a question arises: Why, if 8-year-olds are spontaneously encoding this cue, do so many of them not spontaneously use it?

Analogous questions can be posed for the weight cue. In another paper, Siegler and Klahr (1982) established that a difference in the tendency to encode the weight cue accounts for a corresponding difference between 3- and 4-year-old children who are able to profit from feedback to make the transition from random responding to Rule 1. Yet, the same 4-year-olds who spontaneously encode weight, do not spontaneously use weight as the basis for their predictions.

To summarize, the readiness studies raise two questions:

1. Why do children of one age spontaneously encode a cue that children at a younger age do not encode? Eight-year-olds spontaneously encode weight and distance; 4- and 5-year-olds spontaneously encode weight but not distance; and 3-year-olds spontaneously encode neither.
2. Why do some children who spontaneously encode a cue fail to use it, whereas others who are just a little older both use and encode the cue?

To my knowledge, no fully adequate answer to these questions was given within the context of a system of rules. Klahr and Siegler (1978) discussed the use of production system models to capture these rules, and they stated that these models can provide adequate descriptions of the state of knoweldge, if they are supplemented by further assumptions about different *encoding operators*. Thus, the difference between the 3-year-old and the 4-year-old is the encode weight operator; the difference between the 4- and 5-year-old is the availability of productions that implement Rule 1; the difference between the 5-year-old and the Rule 1 8-year-old is the encode distance operator; and so on. But little was said in any of these papers about what leads to these differences. The rule approach describes the different states of knowledge, but does little to explain the transitions between these different states.

Siegler (1983) recognized these limitations, and called for increased emphasis on mechanisms of transition. In several recent writings (Siegler, in preparation; Siegler & Munakata, 1993), he suggested that one source of transitions may be change in the probability with which children use different *strategies* (rules and operators, in the earlier terminology of Siegler, 1976, and Klahr & Siegler, 1978). But little was said in what has been written to date about where wholly new rules and operators come from, and it is unattractive to assume that all of developmental change can be adequately understood as a change in probability of selection of pre-existing elements, even if we allow that some of the work will need to be done by combinations of elements, as Siegler and Munakata (1993) suggested.

A CONNECTIONIST APPROACH

The connectionist approach, sketched at the beginning of this chapter, provides a different view of the developmental process. The key difference is that the knowledge underlying performance is not represented in terms of the presence or absence of particular rules, operators, or productions, but in terms of graded connection strengths that may be approximately describable in terms of such symbolic constructs. The approximate descriptions may be useful for providing characterizations of performance (for example, Siegler's Rule 1 is accurate in describing the balance scale performance of many 5–8-year-olds), but do not give insight into the fuzzy edges of performance demonstrated by the torque difference effect or to the developmental progression that underlies the transition from performance characterizable by one rule to performance characterizable by another. Here I show how the connectionist system accounts for much of the same data and provides a way of understanding both the fuzzy edges that we see in many cases and the apparent transitions between discrete states.

Before describing the connectionist system, I stress that there are some findings in the balance scale domain that suggest that the connectionist models do not provide the full story. One example arises in the case of subjects who meet Siegler's criteria for Rule 4, when stringently tested with problem sets like the ones used by Siegler (1976, 1981) that cannot be passed using other compensatory strategies. Data from Wilkening and Anderson (1991), using the functional measurement approach, indicate that most adult subjects use a combination rule that is more additive than multiplicative when adjusting weight or distance on one side of a scale to balance a weight–distance configuration on the other. Assuming (as I do) that this task taps subjects implicit rules rather than explicit strategies, the Wilkening and Anderson data suggest that subjects would not adhere to Rule 4 unless they were actually explicitly multiplying. It is clear from several aspects of

Siegler's data that many of the subjects in his experiments who conform to the Rule 4 pattern actually multiply weight times distance to compute a torque for each side, and then decide which side will go down by comparing the numerical values of these torques through explicit, verbally reportable, arithmetic operations. Among the relevant evidence is the fact that college students and 8th graders can be taught to follow this procedure. Even though few such students spontaneously conform to Rule 4, they can come to do so if given an explicit record of the problems or hints to formulate an explicit rule that considers the number of weights on each side, and their distances from the fulcrum (Siegler & Klahr, 1982). This is not to say that successful navigation of many sets of conflict problems requires explicit use of Rule 4; some sets of such problems can be solved by additive or mixed combinations of weight and distance of the kind I believe characterize intuitive judgments. My claim is that subjects' implicit judgments do not closely mimic a strict multiplicative integration rule, and in cases where great care has been taken to make it difficult to succeed using anything other than strict multiplication of weight times distance, few subjects succeed unless they do use explicit multiplication. People can and do use explicit strategies in some tasks and under some circumstances, and the balance scale task is one that appears to elicit explicit strategies under some conditions.

The main interest of this chapter is in the earlier stages of development that lead up to the Rule 3 stage, where the subject takes both weight and distance into account, but does not know explicitly how to combine them to perform at the Rule 4 level. I claim that performance up to this stage (which characterizes most adults, unless specific emphasis and coaching is given, leading to discovering and articulating the rule) can be based on implicit, graded (connectionist) knowledge, and progress through the stages is based on implicit, incremental learning. There is a role for (conscious, explicit) symbolic rules. In the discussion at the end of this chapter, I examine the role such rules might play and consider how they might interact with connectionist forms of knowledge representation.

The Connectionist Model of McClelland (1989)

The connectionist model is based on the learning principle previously stated: The model is trained on examples of balance scale problems. First, the problem is presented (some number of weights on each side of the scale, placed some distance from the fulcrum on each side). The model must try to predict which side will go down. After the prediction, the network is given feedback in the form of the correct outcome for the problem. Then, the weights are adjusted in accordance with the principle previously stated: Adjust each weight in proportion to the extent that its adjustment will

reduce the discrepancy between the model's output (the prediction) and the observed output (the correct response).

To turn this abstract principle into an explicit model, we must make several additional stipulations. First, we must specify a format for representing the problem, both for the input and the outcome. Second, we must specify a network architecture in terms of units and their activation function. Third, we must specify a training regime. The bulk of the work reported here is based on the approach I used in earlier simulations (McClelland, 1989), but in a later section of the chapter, I consider an alternative approach.

In my 1989 work, following up on an earlier model by Jenkins (1989), I chose a way of representing the information needed to solve the problem that was, on the one hand sufficient to distinguish the different possible problem configurations but that, on the other hand, left the network with a substantial task to solve in determining how to interpret the weight and distance information (Fig. 4.5). To allow the network to handle problems involving one to five wieghts on pegs spaced one to five steps from the fulcrum on either side, I provided a total of 20 input units, one to represent

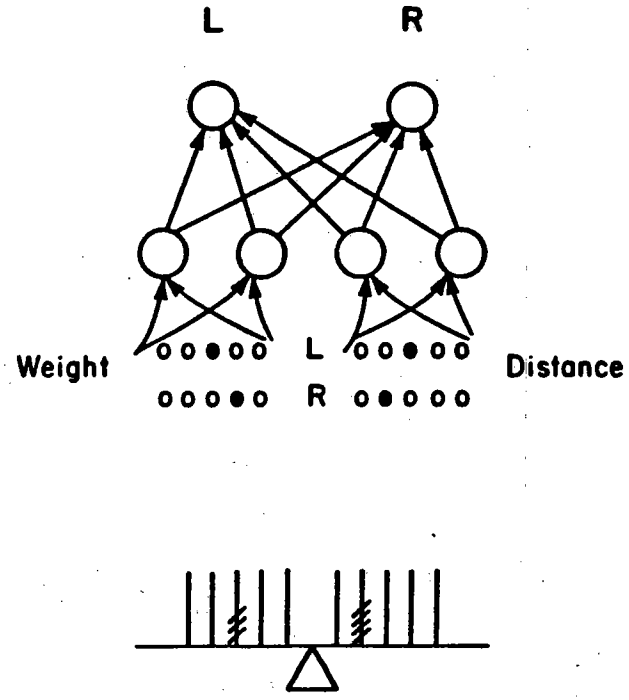

FIG. 4.5. Connectionist network used in the McClelland (1989) balance scale model. Reprinted from Figure 2.7 of McClelland (1989), with permission.

each of the distinct numbers of weights (1–5) on each side, and one to represent each of the distinct distances from the fulcrum of each side. In the figure, the units are arranged into four groups of five, with the two groups on the left representing weight, and the two groups on the right representing distance. With this input representation, a problem can be presented by turning on just the four input units representing the number of weights on each side and their distances from the fulcrum. Note, though, that the input representation only treats the different numbers of weights and the different distances as cardinal numbers; although the units are arranged in increasing order for our convenience, the network has no access to this arrangement, and as far as it is concerned they could be arranged in any other way. At the output level, there were two units. The outcome *left side down* is represented by an activation of 1 on the left unit and 0 on the right; *right side down* is represented by 0 on the left and 1 on the right; and balance is represented by an activation of .5 on both output units. This choice does constrain the network to treat balance as intermediate between the two other alternatives, and injects prior knoweldge of the semantics of the domain into the network.

The network architecture had two other important features: First, it introduced a layer of hidden unit between input and output; and second, it organized these into separate modules, one for encoding the weight information and one for encoding distance. The two-layer structure of the network was imposed to capture the idea that the child must do two things with the information about each dimension: encode that information, and then use the information to predict which side will go down. To be sure, the weight and distance information are encoded in the input to the model. But we can treat this input as corresponding as far as the model is concerned to something akin to the percept in children. Surely, even the youngest children in any of the studies we are considering see—in some sense encode—both the magnitude of the weight (or at least the height of the stack of weights) and its location within the balance scale. We could demonstate this by asking them to point to the top of the stack of weights on each side of the scale. The input representation is intended to capture this level of encoding. But Siegler's (1976, 1981) studies suggest that children differ in the extent to which they encode the relevant dimensions in a form that makes them suitable for predicting the outcome of the balance scale or even for reproducing this information in a copy of a presented balance scale configuration. The intermediate layer of units in the model provides a level that will correspond to this recoding of the perceptual information. The modular organization was imposed to constrain the kinds of solutions the model can find, but as I discuss later, work by Schmidt and Shultz (1991) suggests that imposing this constraint is not crucial.

So far, I have not provided the model with any basis for earlier mastery of the use of weight as a cue as opposed to distance. A definitive treatment

of this issue will require a fuller psychological investigation. It is not immediately clear why the weight cue is noticed and used at an earlier age than the distance cue. One possibility is that children have more relevant experiences with variations in weight than they have with variations in distance. It is a widely accepted principle of language acquisition that children learn to use first those cues that are most available as predictors of the correct interpretation (Bates & MacWhinney, 1987). It is likely that the same principle holds in other domains, as well, and the earlier mastery of weight as opposed to distance may be a case in point. One possible relevant source of experience is see-saws, because see-saws are generally set up with a seat equidistant from the fulcrum on either side. Thus, every child will have had experience with the effects of weight differences, but they may have had considerably less experience with effects of differences in distance from the fulcrum. In accordance with this possiblity, the training regime used in the McClelland (1989) simulations involved presenting the network with training cases that contained many more instances of problems where weight varied but distance stayed the same than of any other type. The exact training regimen consisted of creating a corpus of examples consisting of all possible combinations of one of five weights with one of five distances on the left and the right. This yielded 625 distinct problems. The list was augmented with additional copies of each problem involving weights placed the same distance from the fulcrum on both sides. In two runs, there were five copies of each problem of this type; in two other runs there were ten.

In each run, the network was initialized with small, random connection weights. A series of training epochs was then constructed. In each epoch, 100 patterns were chosen at random from the corpus just described. After the presentation of each pattern, the correct answer was presented, and the weights were adjusted a small amount according to the gradient descent learning rule (see McClelland, 1989, for further details). At the end of each epoch, the network was tested on a set of 24 problems modeled after the 24-problem test set used in Siegler (1981), including 4 problems of each type. On each problem, the activation of the two outputs units was compared. If they were within .33 of each other, then response was taken to be balance; otherwise, the network was taken to have predicted that the side corresponding to the unit with the greater activation should go down. This thresholding corresponds to an assumption that the discrete responses in Siegler's task actually reflected an underlying continuity in the internal psychological states.

Basic Simulation Results

The simulation results were presented in McClelland (1989), so I give a brief summary of the main points so that we can focus on some details not

emphasized in that paper, and on new simulations. First, the model's performance corresponded to one of Seigler's rules about 85% of the time, not counting an initial phase of about 10 epochs. This 85% conformity figure is less than is typically found in subjects, though not by much. Second, the model exhibits good fidelity to the developmental trends seen in human subjects: All four runs of the model showed a plausible developmental progression, starting with no rule at all, and progressing through relatively stable performance on Rule 1 to Rule 2, to a period of vacillation between fitting the criteria for Rule 3 and Rule 4. Thus, the model captures the expected developmental progression of children, at least at a coarse grain of analysis.

An examination of the connection weights in the model provides some

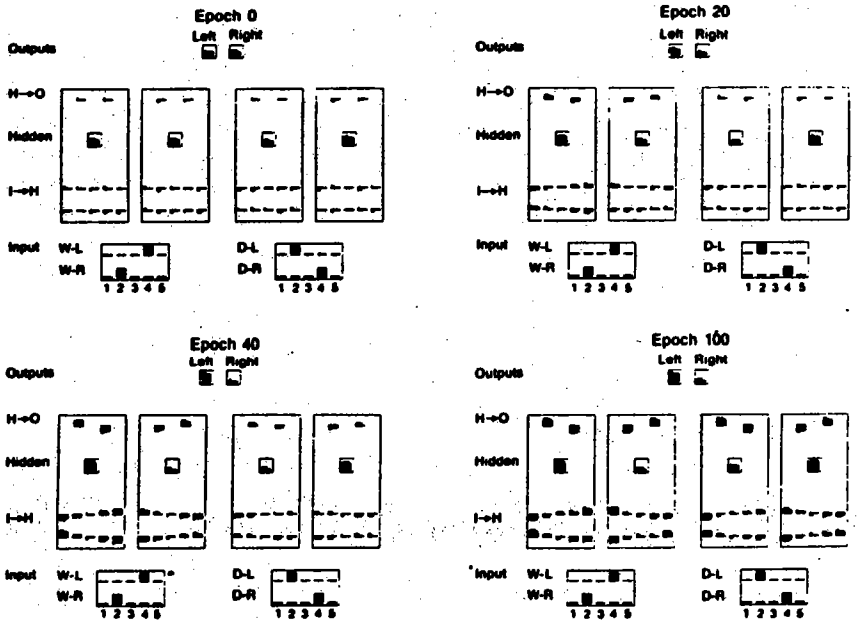

FIG. 4.6. Connection strengths in the McClelland (1989) balance scale network after different amounts of training, and activations produced by an input consisting of four weights two pegs from the fulcrum on the left and two weights four pegs from the fulcrum on the right. The inputs are shown in the lower part of each of the four panels of the figure. The four large rectangles in each panel display the input and output connections and the activations of each of the four hidden units; activations of the two output units are shown at the top. Activations range from 0 (a horizontal line with nothing over it) to 1.0 (a horizontal line with a solid square block over it), with intermediate values represented by partial blocks. Connection weights range from about −6 to about +6; the sign of each weight is indicated by placing the block above or below the horizontal line, and the magnitude is represented by the size of the block. Reprinted from Figure 2.11 of McClelland (1989), with permission.

insight into how it has learned to perform the balance scale task. Figure 4.6 shows the weights in the network from one run, tested before the beginning of training and then after 20, 40, and 100 epochs. For this run, these points correspond to the beginning of the Rule 1 phase; the end of this phase as the model is about to progress to Rule 2; and the end of the training regime, where it is vaccillating between conformity to Rules 3 and 4.

Initially, the model knows nothing about the task. The connection strengths are small and random and the activations of both output units are near .5 for all inputs. (In this part of the chapter, I use the word *strength* to refer to the magnitude of connections; the word *weight* is reserved for use in reference to the weight variable in the balance scale problems.) Gradually, over the course of training, the connection strengths become organized. We will consider the time course in more detail later. For now, we note that by 20 epochs, this organization has begun for the connections that process the weight cue; but at this point, the network has not reached the point where there is any discernable organization in the connections that process distance. This difference is due to the fact that the network receives considerably more experience with cases in which weight varies as a cue than with cases in which distance varies.

Let us examine the way in which the network has leared to encode the weight dimension. The two representational units have organized themselves so that the activation of one ranges from 0 to 1, as the disparity in number of weights between the two sides (number of weights on the left minus number of weights on the right) varies from −4 to +4; the other unit varies from 0 to 1 as the weight disparity varies from +4 to −4. This happens as a result of the connection strengths that the network has learned from the input to the hidden units. These place the different numbers of weights on a continuum, with 1 weight and 5 weights having relatively extreme influences (excitatory or inhibitory), and 3 weights having nearly no influence. In the case of the input given in the figure (weight disparity of 2) the first unit takes an activation of about .7, and the second of about .3. In short, the connections from the input units to the hidden units encode the disparity of weight between the two sides of the scale. The connections from these units to the response units then implement the rule: *The side with the greater weight goes down.* The hidden unit whose activation increases as the weight disparity varies from −4 to +4 tends to excite the left output unit and inhibit the right output unit, whereas the other hidden unit has opposite influences. Thus, when the disparity is 0, the influences on the output units are in balance, but when the disparity favors one side or the other, this is reflected in a disparity in the activation of the output units. At this point, the network has just reached the point where its behavior corresponds to Rule 1.

Considering the connection strengths at 40 epochs, we see that the pattern in the connections encoding the number of weights has become stronger, and a similar pattern has begun to emerge in the connections that encode distance. At this point, this latter pattern is still too weak (and the connection strengths from the hidden units for distance to the output units are still to weak) for the distance information to affect the choice of responses, except in cases of extreme distance disparity and little or no disparity in weight. However, the connections encoding distance and the connections that use this encoding to influence the predicted outcome have built up to the point where slight further increases will now produce sufficient disparities in the activation at the output level to allow the distance cue to influence responses when the weight is the same on both sides. Yet the connections in the distance pathway are too weak to counteract even small disparities in weight. Consequently, the network's behavior will conform to Rule 2 after these slight further increases. Disparity in distance influences the behavior of the network only when there is no disparity in weight.

Considering the connection weights at 100 epochs, we see that they are equally strong on both sides of the network. This may seem surprising, because the network experiences far more cases where the weights differ but the distances are the same, than cases in which the distances are the same throughout training. The reason things level off eventually is that they reach a point where further increases in connection strength on the weight dimension do not lead to further improvements in performance. At this point, the network's output does not match the exact target values, but the slight changes to the connection strengths that occur as a result of these discrepancies cancel out, because they help with some patterns, but not with others. Meanwhile, there is continuing improvement in encoding and use of information about the distance cue. At this point in training, there is still a slight advantage for the weight dimension relative to the distance dimension, which is why the output shown in the figure shows a slight difference favoring the side where there are more weights. The exact outcome depends on the exact input configuration. The network tends to prefer the side with more weight to the side where the weights are a greater distance from the fulcrum, but there can still be cases in which this slight preference is reversed or neutralized, so the network's behavior over a set of problems is somewhat unsystematic, and therefore appears to conform to Siegler's Rule 3.

Sources of Variability in the Network's Behavior

The network can, if trained very gradually, find a set of connection weights that will allow it to perfectly simulate Rule 4 for the range of values of weight and distance used in these simulations. This behavior is fragile,

however, and depends on exact values of connection strengths that are difficult to maintain. The variability introduced by the random sequence to training trails tends to disrupt Rule 4-like performance. In other studies (McClelland, 1991, 1993), I argued that human information processing is intrinsically variable or noisy. Although the present model is deterministic in the sense that the output it generates is a deterministic function of the input and the connection strengths, there is one source of variability from epoch to epoch, namely the random sample of training examples. This, together with use of a learning rate constant large enough so that the random sample of training cases in each epoch exerts a marked enough effect on the connections, introduces fluctuations in the connection strengths from epoch to epoch that produce some inconsistency. This inconsistancy affects different problems of the same type as well as performance on the same problem when the network is tested after different epochs. Although a fully realistic model would, I believe, incorporate processing variability (McClelland, 1993), the variability introduced by the random sequence of training trials has similar effects. Note that the variablity actually affects the activations of the units in the network for all problems, but only affects the actual overt choice of response on some problems. The activations have more of a margin for error in some cases, so they are robust against the amount of variability that arises from the training regime used in this model. The presence of this variability does prevent the model from strictly capturing Rule 4, but this is as it should be, given the evidence previously discussed that exact adherence to Rule 4 depends on use of an explicit multiplicition of numerical quantities.

Discrepancies Between the Connectionist Model and Use of Siegler's Rules

Thus far, we have seen how the model can be used to produce behavior that conforms to Rules 1, 2, and 3, and can occasionally approximately conform to Rule 4, depending on the pattern of connection strengths. If we look at a finer grain, we see that the model does not conform exactly to the pattern of responses predicted by any of the rules. In fact, the model and human subjects deviate from the rules in similar ways, as can be seen by comparing the pattern of performance exhibited by the children and the model to the patterns predicted by Siegler's Rules, in Fig. 4.4. For example, we see in the first panel of this figure that when the model conforms to Rule 1 by Siegler's criteria, it nevertheless occasionally fails on balance problems and occasionally succeeds on distance problems; a similar pattern occurs with human subjects. Similarly, with Rule 2, the model and subjects both tend to make errors on balance and distance problems, and occasionally to get conflict-distance problems correct. With Rule 3, both the model and human subjects

make more correct responses on conflict-weight problems than on other types of conflict problems. Finally, with Rule 4, both the model and human subjects err on the conflict problems, not on problems of other types.

There is one place where the model's behavior deviates from the rules in a way that children's behavior does not. This occurs on conflict-balance problems when overall performance conforms to Rule 3. There is a tendency in the same direction under Rule 4, and in a recent replication with a refined version of the 24-item rule-assessment test, a similar deviation occurs under Rule 2. The discrepancy is due to the fact that human subjects tend not to use the balance response unless the scale is symmetrical (same number of weights, same distance from fulcrum, as in the case of balance problems). When it is asymmetrical, as in all types of conflict problems, they appear to adopt a relatively stringent criterion for the balance response. The model does not have the perceptual capacity to detect symmetry, so it cannot apply a different criterion to the two different kinds of cases. I leave it to further research to explore incorporating some form of symmetry detection capability into the model.

For the cases where the model and the children differ from Siegler's rules in similar ways, we might ask what causes these differences. In the case of the model, they arise from two sources. First, the variablity previously discussed introduces some errors; as already stated, these tend to come at places where correct performance depends on exact numerical values of connection strengths. In addition, some discrepancies also arise from the continuous nature of the model's representations. This means, for example, that large discrepancies tend to produce bigger activation differences than smaller ones, as in algebraic models such as the one described by Equations 1 and 2. Because of the graded connection weights, the connection strengths in one of the pathways can be strong enough to make a difference in extreme cases, but not strong enough to do so in subtler cases. So, for example, just before the model makes the transition to full conformity to Rule 2, it performs correctly on extreme disparities in distance but not on small disparities. Similarly, near the end of training, the model can get some conflict-distance problems right if the disparity on distance is much greater than the (conflicting) disparity in weight; and it can get some conflict-weight problems wrong (giving occasionally the balance response) when the disparity in weight only slightly outweighs the disparity in distance. This occasionally leads to cases where the model approximately corresponds to Rule 4, even though it is not in strict correspondence on every problem.

Overlay of Explicit Rules?

The fact that children deviate from Siegler's rules in most of the same places as the model provides further support for the idea that for children there

is an underlying continuous representation that varies in strength and is best thought of as approximately captured by Siegler's rules. However, although the discrepancies from the rules tend to occur in the same places in the model as in children, the model deviates more strongly from the rules than children do in almost every case. Indeed, quite often the children fall close to the midpoint between the model's performance and the performance predicted by Siegler's rules. The question arises, then, why it is that the children's behavior comes closer to Siegler's rules than the model's does. One answer may lie in the possibility that children do sometimes use such rules. Perhaps the patterns we see with children reflect a mixture of cases of explicit use of rules combined with other cases in which an implicit, activation-based strategy is used.

Children's tendency to use explicit rules may vary with details of the experimental situation. Some aspects of Siegler's methods may have influenced some children's tendency to use explicit rules in his experiments. In Siegler (1976), some subgroups of subjects in Experiment 1 were explicitly instructed to try to discover the rule that governs the behavior of the balance scale. In Siegler (1981), all subjects were run on three different problems with the same formal structure as the balance scale task, and after each task, each subject was asked to describe how he or she solved the problems on which he or she had just been tested. Because order of tasks was counterbalanced, two thirds of the subjects in the balance scale task would have the expectation that they would be asked to explain the basis of their categorization performance. The need to produce an explicit verbal statement of the procedure they used may have influenced subjects to formulate and use such rules in both of these experiments. Interestingly, the likelihood that a subject's behavior would conform to one of Siegler's rules was higher in his experiments than in experiments by Ferretti et al. (1985) or Ferretti and Butterfield (1986), in which subjects were tested in groups, used paper and pencil to indicate their response, and were not asked to describe how they solved the problems.

Even in cases where there is no explicit expectation that they will have to explain their behavior, subjects may sometimes formulate explicit rules; Karmiloff-Smith (1986) argued that childen may have a natural tendency to try to find a rational basis for their responding in the form of an explicit rule that seems at least approximately right in that it comforms to their implicit response tendencies. Indeed, they may then use the Rule, as opposed to the implicit response tendency, to actually determine their responses. The fact remains, however, that there are discrepancies between subject's responses and the explanations they give of their own behavior. That these discrepancies often accord with the connectionist model's predictions for where such deviations should occur, supports the view that the rules by themselves do not tell the whole story.

Torque Difference Effects

As previously noted, the torque-difference effect of Ferretti and Butterfield (1986) is another phenomenon that strains the notion that children's balance scale responses can be fully understood in terms of the consistent use of a simple explicit rule like Siegler's Rules 1–4. Connectionist models and others that use graded activations and connection weights provide a natural framework for capturing these phenomena. Indeed, Schmidt and Shultz (1991; Shultz, Mareschal, & Schmidt, in press) have shown that both the McClelland (1989) model and their own connectionist model predict torque difference effects. In this section, I explore whether this sensitivity, at least as exhibited in the McClelland (1989) model, permits a good fit to the Ferretti and Butterfield data.

I consider both the influence of torque difference on probability correct responses to different problem types, and the effects of torque difference on the apparent use of Rules 1–4. To do this, I constructed a new test set for the 1989 balance scale model following as closely as possible the procedure Ferretti and Butterfield used to construct their test. Slight differences arose because Ferretti and Butterfield used a scale with six pegs on each side and up to six weights on a peg, but the McClelland (1989) model used only five pegs on a side and up to five weights on a peg. Thus, for simple weight and distance problems, Ferretti and Butterfield used product differences of 1, 3, 12 and 24–30 for the four different levels; I was able to match the first three values but had to use a smaller range, 16–20, for the fourth level. For conflict-weight and conflict-distance problems, Ferretti and Butterfield used differences of 1, 3, 5, and 18–24 units; again, I was able to match the first three values, but had to use a smaller range, 11–15, for the fourth level. The simulation was run just as in McClelland (1989), with the only differences being: (a) the new Ferretti and Butterfield test was given along with a test based on the one used by Siegler (1976) at the end of every epoch, (b) results were based on eight simulation runs of 70 epochs, rather than the four runs of 100 epochs used in McClelland (1989). The simulation was cut off at 70 epochs because of the young age range of Ferretti and Butterfield's subjects (5–11, compared with 5–20 in Siegler, 1981). Because Ferretti and Butterfield's youngest subjects were first graders who performed at the Rule 1 level in Siegler's test, the results presented were taken only from epochs after the network reached the Rule 1 level—this meant that the earliest 10 to 15 epochs of each run were discarded.

The new simulation captures many aspects of the Ferretti and Butterfield (1986) findings on the effects of torque difference. Consider first, the accuracy data as a function of torque difference (Fig. 4.7). As in the Ferretti and Butterfield study, performance on simple weight problems is highly accurate at all levels of torque difference, and so is relatively unaffected

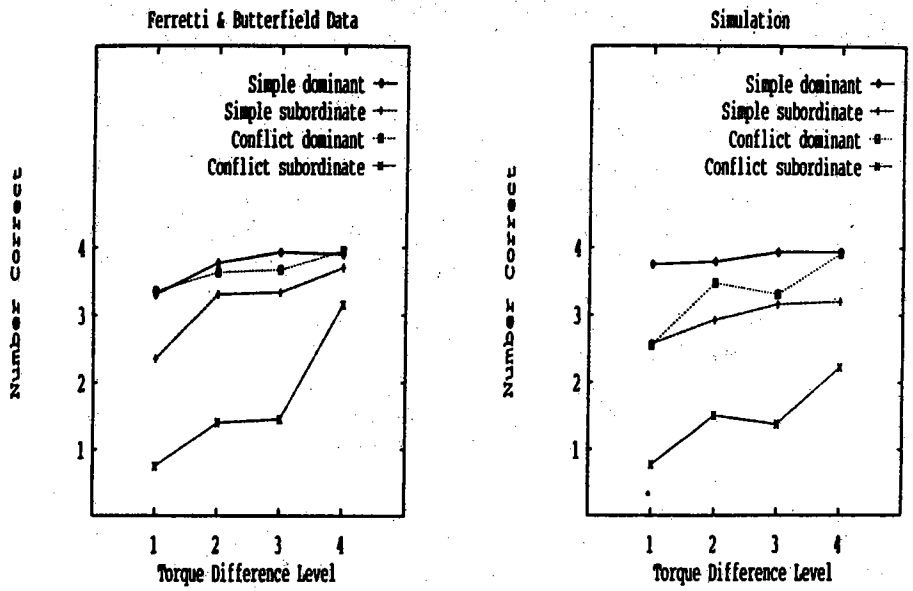

FIG. 4.7. Average number of responses correct on problems of each of four different types, as a function of level of torque difference. Left panel, human subject data redrawn from Ferretti and Butterfield (1986). Right panel, new simulation results based on the McClelland (1989) balance scale network.

by variations in torque difference, but performance on simple distance and conflict-distance problems improves across levels of product difference. The simulation shows a wider variation on conflict-weight problems over torque difference than is seen in the children's data. Also, the simulation shows less improvement than the children on simple distance and conflict--distance problems at the highest torque difference level. This may be due to the fact that the largest torque differences are less extreme in the simulation than in the actual experiment.

Considering the rule classifications shown in Table 4.3, again the patterns seen in the simulation are similar to the patterns seen in children's responses. As in the experiment, the percentage of Rule 1 designations goes down and the percentage of Rule 4 classifications goes up as torque difference goes up, and both the children and the model show inconsistent trends with Rules 2 and 3. The correspondence to the data is good, with the one discrepancy that the simulation shows more Rule 3 classifications at the lowest level of torque difference than the subjects do.

Although the correspondence between the model and the Ferretti and Butterfield (1986) data is not perfect, it is close enough in several respects to suggest that the mechanisms used in the model and the mechanisms used by children have something in common. Ferretti and Butterfield suggested that

TABLE 4.3
Percentage of Rule Classifications at Different Levels of Torque Difference

TD Level	1	2	3	4
1. Ferretti & Butterfield Data				
1	0.29	0.19	0.17	0.05
2	0.24	0.34	0.14	0.08
3	0.22	0.31	0.22	0.10
4	0.19	0.15	0.15	0.37
2. Simulation				
1	0.23	0.16	0.37	0.01
2	0.19	0.15	0.22	0.20
3	0.13	0.22	0.29	0.12
4	0.13	0.18	0.10	0.37

children may use rules but ignore small differences. The model provides an alternative interpretation: It suggests that children may rely on the same kind of graded, activation-based process that underlies the model's performance.

Readiness

We now turn to the findings that are of most interest from the point of view of mechanisms of development, namely the findings on the readiness to progress from stage to stage. In McClelland (1989), I simulated Siegler's (1976) experiments on readiness by examining the effect of training on conflict problems using two different networks. The first was a network just entering the Rule 1 phase, and the other was a network that was just about the exit the Rule 1 phase. These are the networks whose weights are shown at Epoch 20 and Epoch 40 in Fig. 4.6. Each network was trained with 16 examples of conflict problems chosen according to the procedures described by Siegler (1976). Siegler's subjects saw the training examples only once, but just to see what would happen, I presented the 16 examples several times, testing on the 24-item rule assessment test after each presentation of the 16-item training set. Performance profiles for the two networks over problem types, after each set of presentations of the conflict problems, are shown in Fig. 4.8. The results are quite dramatic: The network trained with conflict problems just as it is entering the Rule 1 phase regresses, whereas the network trained with conflict problems just as it is exiting this phase moves forward to Rule 2 after the first exposure to the training problems.

The reasons for these different patterns can be found by examining the connections in the network at the time when the conflict training trials are

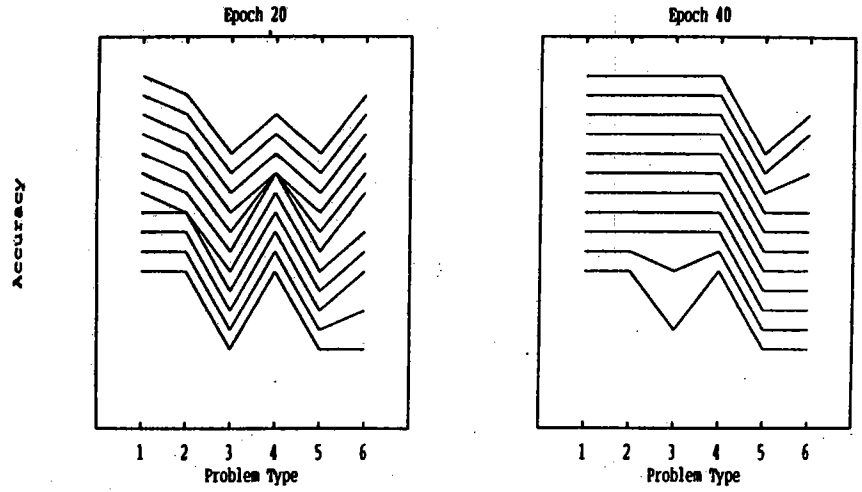

FIG. 4.8. Performance profiles on the rule assessment test after 0 through 10 exposures to a set of conflict problems, after 20 (left) or 40 (right) epochs of training. The figures present successive profiles proceeding back in depth, with the lowest (frontmost) profile in each figure representing performance before the first exposure to the conflict training patterns. Problem type 1 = balance, 2 = weight, 3 = distance, 4 = conflict–weight, 5 = conflict–distance, 6 = conflict–balance as in Figure 4.4. The frontmost profile for the 20-epoch network corresponds exactly to the pure Rule 1 pattern, but with exposure to the conflict problems this profile deteriorates and becomes unclassifiable after the fourth exposure. The frontmost profile for the 40-epoch network corresponds to the Rule 1 pattern with one deviation in the form of one correct response out of four on distance problems. With just one exposure to the conflict problems, the pattern switches over the Rule 2 pattern with one deviation, and even this deviation disappears with just one more exposure.

introduced. In the 20-epoch network considering first the connections from the input units to the internal representation units, there is a definite but still weak (compared to later stages) representation of weight on the two sides of the scale, and virtually no representation of distance. Similarly, considering the connection weights from the internal representation units to the output units, the connections in the pathway that deal with weight encode the fact that the side with the greater weight goes down. For example, the internal representation unit on the left, whose activation varies positively with relative weight to the left, activates the left side down output unit. Again, though, the connections involved are weak at this stage compared to Epochs 40 and 100. Finally, on the distance side of the network, neither unit has significant strength to either output unit. In summary, this network has connections that allow a definite but still weak encoding of relative weight, and a definite but still weak use of this encoding to influence its prediction as to which side will go down. However, it has no encoding of relative distance or of how to use it to influence its predictions as to which side will go down.

Given this situation, consider how the conflict training problems influence the connections in the network. The set of conflict problems present inconsistent feedback with respect to usefulness of the weight cue to predict which side will go down. Indeed, on half of these trials, the side with greater weight goes down; but on the other half, the side with greater weight goes up. Thus, to a network encoding and predicting based on the weight cue, its predictions are in error on one half of the conflict-weight training items. The network makes adjustments to the connections to reduce this error, and in so doing, it reduces the strengths of the connections it uses to encode and use the weight cue. The result in gradual deterioration of performance.

Although the connections are eroding on the weight side of the network, little is happening to the connections on the distance side. Changes to these connections cannot help the network very much. For one thing, as with the weight cue, the feedback is inconsistent. But there is another reason why changes to these connections cannot help. The problem can be seen most easily in considering the connections from the internal representation units to the output units. Changes in these connections do not help performance overall, because the activations of these units do not meaningfully reflect relative distance from the fulcrum on the two sides of the scale. The learning algorithm will increase the strength of the connection from the left-hand distance representation unit to the left side down output unit on a trial where the left side should go down, and will decrease the strength of this connection on a trial where the right side should go down. These influences have no net effect because the activation of the left-hand distance representation unit does not covary with relative distance. So, even if relative distance covaried perfectly with which side goes down, the connections from the distance representation units to the output units would bounce back and forth and little net progress would ensue. A similar problem arises for the connections from the distance input units to the distance hidden units. Given that the connections from the distance representation units to the distance output units are weak, changes in the connection from distance input units to distance representation units have no effect on the predictions of the network. In back propragation, connections are adjusted in proportion to the extent that their adjustment reduces the discrepancy between predicted and observed outcomes, and changes in these connections do not much affect this discrepancy, so the weights are not changed very much.

In summary, conflict training erodes the connections on the weight side of the Epoch 20 network and has little influence on the connections on the distance side. Therefore, this network gradually regresses with repeated presentations of the conflict feedback problems.

Now let us consider the 40-epoch network, also shown in Fig. 4.6. This network differs from the 20-epoch network in two ways. First, the connections on the weight side encode relative weight more strongly and use this cue more forcefully to influence predictions than in the 20-epoch network. Second, the connections on the distance side of the network now encode and use the distance cue, albeit weakly. At this point, the network's profile on the 24-item test shows that it gets one of the four distance problems correct, indicating partial sensitivity to distance cue.

In this case, training with the conflict problems turns out differently. First, consider conflict-weight problems. In these cases, the weight cue dominates and there is little error, and therefore, little change to the connections. Now consider conflict-distance problems. In these cases, the weight cue tends to dominate the output, too, and so there is error—the network makes the wrong prediction. On these trials, the connections on the weight side of the network will be eroded somewhat, but because they are stronger than at 20 epochs, this does not readily reduce the network's use of relative weight to predict which side will go down. This time, the connection weights on the distance side of the network are not eroded; they tend to be strengthened. There is now a basis in the existing connection weights on the distance side of the network on which to build. The distance representation units now represent relative distance. For example, the left distance representation unit's activation increases to the extent that the differences between the distances on the two sides is greater on the left. This unit tends to be active on conflict-distance trials when the left side should go down. On these trials, an increase in the weight from this unit to the left side down output unit will substantially reduce the error, as will a negative increment to the connection from this unit to the right side down output unit. Therefore, relatively large changes are made to these connections. In contrast, the right distance representation unit tends to be relatively inactive on conflict distance trials where the left side should go down. An increase in the strength of the connection from this unit to the left side down output unit will, therefore, have a small effect on the error, as will a negative increment to the connection from this unit to the right side down output unit. Therefore, only small changes are made to these connections. The result is that larger changes are made in the right places. Corresponding changes occur on conflict-distance problems where the distance is greater on the right. Between these two kinds of cases, there is a resulting increase in the extent to which the distance representation units exert the correct influences on the output units, thereby increasing the network's sensitivity to differences in distance. Similar influences also occur in the connections between the distance input units and the distance representation units, so both the encoding and use of the distance cue are strengthened.

Accelerations and Decelerations in Developmental Change

I hope that the previous discussion allows an intuitive understanding of the process whereby the network can exhibit differential readiness to learn at different points in its development. But this discussion, posed as it has been in nonmathematical form, may tend to give more of an "all-or-nothing" flavor to the effects of connection changes at different points in time than is really correct. Specifically, in describing the effects of conflict training on the connections on the distance side of the Epoch 20 network, the description given makes it sound as though it would be impossible for the network ever to learn to make use of the distance cue, even with a totally consistent relationship between relative distance and outcome in terms of which side goes down. Indeed, the same would apply just as well to the connections in the weight side of the network: Initially, the network has no meaningful representation of relative weight or of how to use relative weight, and therefore changes to the connections in the weight side of the network will have no beneficial effect. How then does the network ever learn? It seems that we are in the same place we were when considering models based on learning rules. In these models, the transition suddenly happens at some point, with no suggestion of why it happens then or why it takes so long within a stage if a single discrete change (adding a new production) is required to get from A to B. In the connectionist system, although the transition itself is not instantaneous, it seems as though it must at least *begin* to happen suddenly at some point. Otherwise, how would we progress from the initial state of meaningless connections and null or canceling changes, to the state where the connections have started to become meaningful and the changes begin to cumulate? Are we again faced with some sort of "immaculate transition," as Siegler and Munakata (1993) put it?

No, there is no immaculate transition. Recall that the network is initialized, not with 0 connections, but with very small random values. This means that, due to the random initial values, one of the two hidden units in the weight side of the network will be slightly more strongly activated than the other in those cases where the preponderence of the weight is to the right. The effect can be slight and initially not consistent over different instances in which the proponderence of the weight is to the right, but there will inevitably be a slight relative advantage. This advantage does not represent prior knoweldge about the problem; it only represents the fact that in the presence of random initial connections, some units will naturally tend to be more suitable for some tasks than others. The gradient descent learning procedure exploits these random initial differences. Connection changes that initially almost cancel do not quite cancel completely, and gradually, differences build up. As they do, structure emerges. There is no point at which one could say, "now the network is encoding relative weight

and before this it was not," but there are accelerations and decelerations. As structure gradually emerges, changes to particular connections can have large effects on the error. According to the learning rule, larger changes are made to these weights, and there is a double benefit because the changes are larger than they would have been earlier, and at the same time, they have larger effects. Accelerations of this sort, both for the weight cue and for the distance cue, are visible in Fig. 4.9. The acceleration in the connections that process the weight cue mark the onset of the Rule 1 phase, and the acceleration in the connections that process the distance cue mark the transition from the Rule 1 phase on to Rule 2 and then Rule 3.

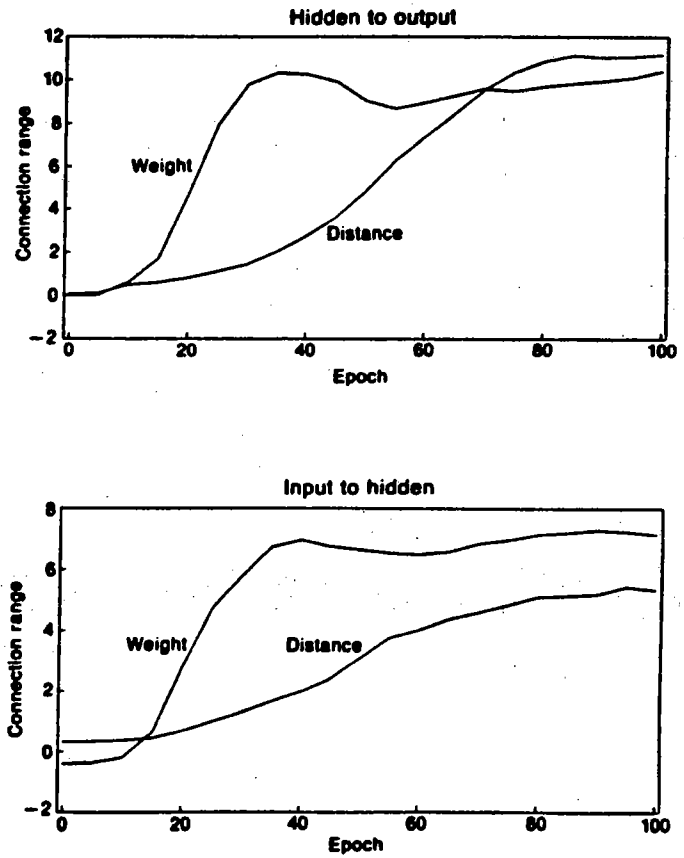

FIG. 4.9. Relative magnitude of connection weights for processing weight and distance as a function of training. Weights for encoding weight and distance (input to hidden units) are shown in the lower panel. Weights for using the results of these encodings to determine which side of the scale will go down are shown in the top panel. Reprinted from Figure 2.12 of McClelland (1989), with permission.

Must We Assume a Biased Environment?

Before turning to a consideration of the implications of connectionist learning for conceptions of developmental change, I would like to reconsider one of the assumptions of the 1989 model. In that model, I assumed that weight is mastered before distance because of differential relative frequency of experience with weight and distance as cues. In terms of the accelerations seen in Fig. 4.9, we see the same phenomenon played out on slightly different time scales, due to the differential frequency of exposure of the model to each of the two dimensions of the problem. Certainly differences in frequency, if they do exist, would tend to influence the time course of developmental change in the use of different cues, but there are other factors that may also be relevant in this and other problem domains.

One such factor is the relative complexity of the cues to weight and distance. Weight is a stable property of an individual object, whereas distance is inherently a relation between a pair of objects. In the case of the balance scale, the second object is the fulcrum. Perhaps the differential difficulty of encoding weight and distance may turn out to depend not on relative frequency but on this difference in complexity of the two cues.

To illustrate this general point, and to show that connectionist models provide a way of addressing it, I carried out one further simulation using the same way of representing weight information as in the earlier simulation, but a different way of representing distance information. Distance is no longer represented explicitly, but instead, the positions of three things—the two weights and the fulcrum—are all represented. A row of units is used to represent the position of the left weight, another row to represent the position of the right weight, and a third row to represent the position of the fulcrum. To ensure that the model actually computes distance, each problem could be presented in any of a number of positions, with the correct response depending not on the absolute positions of the weights but on their relative positions with respect to the fulcrum.

The new network was trained without any bias in the training set at all; each of the 625 different problems that can be made from a random combination of 1–5 weights on either side of the fulcrum at distances 1–5 units from the fulcrum on either side had an equal chance of being presented in each training epoch. Actual location of the weights and the fulcrum varied randomly from trial to trial, so that the network had to take the relative positions of the weights and the fulcrum into account. As before, each epoch consisted of 100 training examples selected at random, followed by rule assessment using Siegler's test. The results, in the form of the best fitting rule at each epoch, are shown in Fig. 4.10. The results demonstrate an obvious advantage for weight over distance as a cue, in that the model exhibits a very extended Rule 1 phase in which distance has no

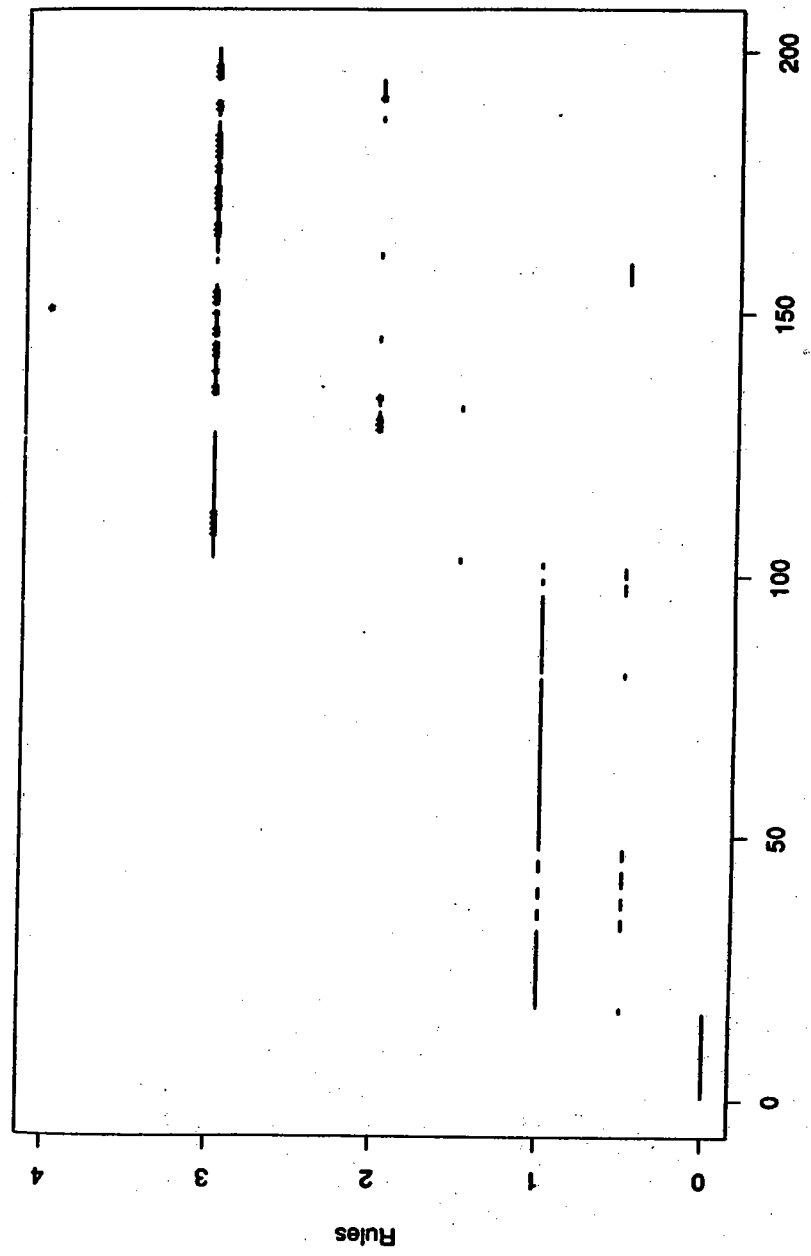

FIG. 4.10. Best fitting rule after each epoch in the new network in which the distance to the weight on each side must be computed based on information about the position of the fulcrum and of the weights on each side of the balance scale. When a dash is used, the best fitting rule fits well enough to satisfy Seigler's (1981) scoring criteria; when a "*" is used there are too many deviations from the ideal pattern to satisfy the criteria, and so the pattern would be unclassifiable.

impact on performance. The results do not provide a perfect fit to all aspects of the developmental data, in that the model never exhibits Rule 2 and there is considerable inconsistency in its performance once it exits the Rule 1 phase. But the results do indicate that connectionist models can easily capture differences in cue complexity, and illustrates how these differences provide another possible basis for developmental differences in the use of different cues.

GENERAL DISCUSSION

This chapter has reviewed earlier work (McClelland, 1989) demonstrating the applicability of a connectionist approach to the apparently rulelike progression of behavior seen in tasks like the balance scale task, and has extended this work in several ways. This work is part of a growing body of connectionist work examining various aspects of development, including research on language development (Elman, 1991; MacWhinney, Leinbach, Taraban, & McDonald, 1989; Plunkett & Marchman, 1989; Rumelhart & McClelland, 1987), conceptual development (Chauvin, 1989, Schyns, 1991), seriation (Mareschal & Shultz, 1993), and other models of performance in the balance scale task (Shultz, Mareschal, & Schmidt, in press). Taken together, these papers provide a growing body of simulation studies suggesting that development may not be a matter of changing systems of rules, but of changing systems of knowledge stored in the form of graded connections. Drawing on these studies, both Karmiloff-Smith (1992) and Halford (1993) have proposed general frameworks for cognitive development that incorporate connectionist principles as the substrate for acquisition of implicit knowledge. Both authors argued that implicit knowledge is only one form of children's knowledge. As I discuss later, I share with these authors the view that implicit knowledge coexists with more explicit knowledge—knowledge that is often verbally reportable and that often corresponds to a system of explicit rules and propositions.

The set of simulations reported in this chapter, taken together with insights arising from the other work just cited, make several important points about the representations that may underlie implicit, intuitive knowledge and the mechanisms through whgich these representations change in response to experience. I summarize the main points arising from the present simulations, discuss how they relate to basic characteristics of connectionist models, and consider some of the broader implications for understanding aspects of cognition and cognitive development.

Sensitivity to Cue Magnitude

First, the simulations provide an account for the sensitivity of performace in tasks like the balance scale task to the magnitudes of the various cues—

in the balance scale task, the magnitude of the disparity in weight and/or distance from the fulcrum on the two sides of the balance scale. These effects, clearly documented by Ferretti et al. (1985) and by Ferretti and Butterfield (1986), as well as other evidence fo the use of algebraic rules in some variants of the balance scale task (Wilkening & Anderson, 1991), indicate that simple rules of the form first proposed by Siegler do not fully characterize children's knowledge of the balance scale in all cases.

This aspect of the model's performance derives from its use of graded connections, rather than all or nothing rules, to capture its relevant knowledge. Of course, this assumption is one the connectionist approach shares with other approaches, such as the *cognitive algebra* approach of Anderson (1991), and other approaches to knowledge representation based on fuzzy logic. It is an assumption that is helpful in providing an account of certain important aspects of the relevant developmental data.

Stagelike Developmental Progressions

Secondly, the simulations account for the general form of the developmental progression seen in the balance scale and many other similar domains. Most importantly, these simulations have shown how stagelike progressions can emerge from incremental, continual change. In this respect, the simulations can be seen as providing an explicit mechanistic basis for one of the main tenets of Piagetian developmental theory. Piaget believed that the process of equilibration produced small changes that accumulated to yield apparently qualitative developmental change, but never showed in detail how this might occur (see Flavell, 1963, for a discussion). The balance scale model reviewed here illustrates how such qualitative changes—apparent stagelike progression—could arise from the accumulation of small incremental changes. In the model, acquisition of the use of each of the two cues begins with an initial phase in which the effects of experience accumulate gradually, followed by a more rapid acceleration. This basic property of connectionist networks, coupled with some basis for easier mastery of one cue than another, is sufficient to account for the developmental progression from initial failure to use either cue, to virtually exclusive reliance on one cue, to a transitional period where both cues are taken into account but one strongly dominates, to a final phase that can be characterized by an approximately additive combination of cues. In other work, examples of this kind of progression toward a fuller implicit grasp of the structure of the domain can be seen. For example, Cleeremans, Servan-Schreiber, and McClelland (1989) trained a simple recurrent connectionist network to anticipate the next element of a sequence based on prior elements. The model shows what I call *progressive penetration* into the structure of the sequence (borrowing a felicitious phrase from Flavell,

1963), in that its performance first corresponds to optimal predictions based only on the preceding element, then progresses successively through phases in which it takes successively more and more elements into account.

Differential Readiness

Third, the simulations account for differential readiness to profit from experience as a function of current developmental state. Although rule-based approaches (Klahr & Siegler, 1978; Siegler, 1976) can characterize the differences as well, they do so in a way that begs the fundamental developmental question, namely, how does the child get to the state of readiness from the state of unreadiness? As previously noted, this is described by Siegler and Munakata (1993) as the problem of the immaculate transition.

These aspects of the model's performance derive from what Elman (1991) called the incremental character of the learning process in connectionist networks. As in Piaget's conception of development, each new accommodation and assimilation builds upon the accumulated residue of past accommodations and assimilations. I tried to make clear how this process works in some detail, both in describing how accelerations occur and in discussing why the network progresses from conflict training at one point in development but not at an earlier point. The essential observation is that the error-correcting learning process relies on the propagation of signals in much the same way that the forward-going activation process does. When connections in the network are random, little is propagated in either direction; but once the connections have begun to capture the structure in the domain, both activation and error signals can propagate effectively.

It has often been observed that back propagation learning is biologically implausible (e.g., Grossberg, 1987). The implication is that models that rely on it use a method that is computationally infeasible for the nervous system, and therefore of little relevance to understanding human cognitive processes. I believe this argument is shortsighted. It is now understood that the computational equivalent of back propagation of error can be implemented in connectionist networks using only the bidirectional propagation of activation (Hinton, 1989; Hinton & McClelland, 1988; Peterson, 1991). As with back propagation, the propagation of information both forward and back depends on the knowledge stored in the connections.

Basis for Differential Use of Cues

Finally, the simulations provide a demonstration of the point that connectionist models provide a natural framework within which to explore the possible basis of developmental differences in the representation and use of

various cues. In a wide range of domains, children tend to focus on one cue when there are two or more that are of equal importance from the point of view of the physical processes in play in the domain. Even if we accept the frequently held view that the tendency to focus on a single dimension across a wide range of tasks reflects capacity limits, one of two dimensions will reliably dominate the other, supporting the notion that there are differences in the strength of the child's sensitivity to the dimensions. The connectionist approach allows us to see how both frequency of exposure and complexity of a cue could play a central role in determining which cue will dominate. Other researchers, particularly Bates and MacWhinney, have stressed the importance of cue frequency (what they call *availability*) in development, and MacWhinney et al. (1989) and MacWhinney and Leinbach (1991) have noted that connectionist models capture the effects of availability (as well as validity and conflict validity) in development of use of articles and inflections during language acquisition.

The present simulations are the first to stress the possible developmental implications of the role of cue complexity in multilayer connectionist networks. In fact, the factor I call *complexity*—specifically, the need to take account of higher order relational information to accurately perform some task—is understood to be of central impoprtance in connectionist networks. Perceptrons (networks with one layer of modifiable weights) are not capable of learning to exploit higher order relational information (Minsky & Papert, 1969). Although multilayer perceptrons trained with back propagation can learn to use relational cues, they learn lower order relationships much more easily; the simulation I reported here (of the relative difficulty of learning to extract the higher order distance cue from position information) is one of many cases in point. Although all the reasons why networks have an easier time with simpler cues have not been fully determined, several contributing factors can be identified. Perhaps the fundamental one is the fact that correlations between input and output become contingent on other input variables. The problem, as the network sees it, is one of learning several such conditional contingencies instead of just one.

To make this concrete, consider a network that is learning a relationship between two binary input variables (x_1, x_2) and a single output variable (y), and consider two relationships: $y = x_1$, and $y = (x_1? = x_2)$. (The second function has value 1 if x_1 and x_2 are identical, value 0 otherwise.) The truth tables for these two relationships are shown in Table 4.4. There is a simple noncontingent correlation between x_1 and y in the first case, but no correlation between either x_1 or x_2 and y in the second case. However, in the second case, there is a correlation between x and y that changes sign depending on the value of x_2. A back propagation network of the kind used in this chapter essentially handles this relation in just this sort of way.

TABLE 4.4
Truth Tables for Two Relations

$y = x_1$			$y = (x_1 ? = x_2)$		
x_1	x_2	y	x_1	x_2	y
0	0	0	0	0	1
0	1	0	0	1	0
1	0	1	1	0	0
1	1	1	1	1	1

Different runs with different random starting connections find slightly different specific solutions, but they all involve at least one hidden unit that learns a contingent correlation. In one typical solution, the network that results from training has one hidden unit that handles the cases where $x_2 = 1$, in which case there is a positive correlation between x_1 and y, and it has another unit that handles cases in which $x_2 = 0$, in which case there is a negative correlation betwen x_1 and y. The learning is much slower in this case than in the case of the simple function $y = x_1$, for several reasons. One is that each of the contingent correlations is only exhibited by a subset of the training cases. A second problem is that the units must learn to specialize; at first, both units are affected by all of the cases and so (in a process we saw in action at other places in this chapter) the changes to the connection weights nearly cancel each other out. As a result, there is a long initial phase where little learning takes place in the relational case.

Halford (1993; Halford et al., in press) has stressed the importance of complexity, measured in terms of the number of interacting problem dimensions that must be considered for adequate performance. Halford proposed the use of Smolensky's (1990) tensor networks to capture complex higher order relations, and specifically proposed a mapping of the balance scale problem onto the tensor product representation. In these representations, each problem dimension is represented by a pattern of activation consisting of a set of elements. Combinations of elements from different dimensions are represented by assigning one connectionist unit to represent every possible N-tuple consisting of one element from each dimension. Obviously on this approach, higher order relations require extremely complex representations. Tensor product representations were used in a wide range of connectionist models (e.g., Rumelhart & McClelland, 1986; Touretzky & Hinton, 1988) but they suffered from combinatorial explosion and from a dispersion of the regularities that must be learned for mastery of the content of particular domains (Plaut & McClelland, 1993; St. John & McClelland, 1987).

The preceding discussion suggests that the use of tensor product representations may not be necessary to capture within a connectionist framework at least some aspects of the empirical findings that suggest that higher order relations are harder to learn. Tensor product networks provide a full set of units, one for every possible conjunction of one element from each dimension. Back propagation networks instead provide a smaller number of units that are assigned by the learning process to capture conjunctions that are necessary to perform a particular task. In both cases, there is an inherent capacity to capture higher order relations, and in both cases, higher order relations are more difficult to capture than simpler ones are.

Is the Feedforward Nature of the Present Model an Inherent Limitation?

Another reason why Halford proposed tensor product representations is the fact that they provide a framework within which knowledge can be flexibly accessed. Tensor product representations treat all dimensions of the problem equivalently, so that any can be used as either input or output. Thus, for example, a tensor product model of the balance scale having separate dimensions for each of the input variables and another dimension for the outcome would allow one to specify all but one input and the outcome, and the model would fill in the missing values. Given the complexities of using tensor product representations, this same flexibility should be possible in connectionist networks that are slight variations on the network used in the simulations reported here. Networks with bidirectional connections can be trained with a variety of learning rules, and there are several extant examples of networks that can take inputs from a subset of the input dimensions and complete the remaining dimensions (given that the subset of dimensions provided in the input provides sufficient constraint on the unspecified dimensions; Hinton, 1981; Movellan & McClelland, in press).

Open Questions for Connectionist Approaches to Cognitive Development

Several fundamental questions remain open at this stage. In concluding, I consider two: The nature of the initial structure that must be built into a network to allow it to learn and to account for the developmental patterns seen in a particular domain, and the relation between what is learned in connectionist networks and explicit knowledge that can be articulated in verbal form.

Nature of the Initial Structure. The McClelland (1989) network built in considerable initial structure by assuming separate input representations for weight and distance cues and separate internal representation units for

learning to represent each of these two cues. To what extent do the results reported here depend on these assumptions? Work by Schmidt and Shultz (1991) cast considerable light on these matters. In one study, Schmidt and Shultz examined the adequacy of a number of variants of my 1989 model, in terms of the extent to which the model would exhibit (a) a high probability of adhering to one of Siegler's four rules after each epoch of training (excluding a few initial epochs), and (b) sequential progression through all four rules. They varied the learning rate, whether or not the hidden units were presegregated, and the degree of bias in the training set favoring exposure to problems in which weight varied but distance was the same on both sides of the scale. Presegregation of the hidden units was not a major factor in determining the adequacy. Although both the learning rate and the degree of bias mattered (generally, more adequate models were obtained with slower learning rates and greater bias), presegregation of the internal units did not. Furthermore, Shultz et al. (in press) showed that adequate simulations are possible using a variant connectionist network that employs a learning algorithm called *cascade correlation,* in which there are no initial hidden units at all, and hidden units are added one at a time to capture error in the network's predictions that it misses without the hidden units.

Although prestructuring the internals of the network is not necessary, all of the networks under consideration—those used by Schultz's group, and those used in my simulations—do assume that the input and output representations are highly structured. It is hardly the case that any of the networks see raw, unencoded perceptual input and extract from this weight and distance. I imagine that the perceptual processes that extract the visual concomitants of these variables from the raw perceptual input draw on hard-wired perceptual mechanisms as well as experience-based learning processes. It is not clear exactly what form these perceptual representations take or what form they need to take to be useful in systems that gradually learn from experience. This is a matter that could be fruitfully explored in future research. Shultz et al. represented different amounts of weight or distance on each side of the balance scale by different degrees of activation of separate units for each weight and each distance, and obtained good simulations of the main phenomena, but other than this variation and the new representation I used in the final simulation reported here, there has not been much work on what aspects of the input (or output) representations are crucial. Input and output representations make a big difference in some cases (Plaut & McClelland, 1993), so it will be important to explore this matter further. The simulations I have reported here say little about the extent of prior constraints that must be built in to allow development to successfully proceed. This is obviously an important issue for future research, but it is not the burden of the present work to address this issue. Rather, the point is to show how experience may be that engine that drives development, through channels determined by initial structure and by the nature of the input and output representations that are used.

Relation of Implicit Connectionist Knowledge to Explicit and Implicit Rules. One of the cornerstones of Siegler's rule assessment approach is the finding that there is a fairly close correspondence between children's verbal reports about how they do the balance scale task and their performance as measured by the rule assessment method. All of the balance scale models, and indeed all of the other extant connectionist models of other developmental phenomena, have nothing to say about this explicit knowledge. Therefore, it is worth considering the status of these explicit reports and their relevance to accounting for subject's behavior.

First of all, both adults and children can and do use explicit rules to govern their behavior some of the time. The connectionist models I have explored here, as well as most other connectionist models, leave out such explicit rules and are obviously missing an important aspect of human cognition. Within the balance scale task, some children use very explicit procedures, at least some of the time. There is good evidence from a variety of sources that older children can and will use Rule 4 if instructed in its use, and can discover this rule for themselves. Although connectionist networks can mimic Rule 4 (or any other deterministic function relating inputs to outputs) arbitrarily closely, implicit knowledge does not generally reach this level of sophistication. In Wilkening and Anderson's functional measurement study, where explicit numerical calculation is not possible due to the use of continuous stimulus and/or response variables, the data indicate that the combination rule is more additive than multiplicative. Thus, subjects who pass Siegler's version of the Rule 4 test, in which the questions were chosen so that the test cannot be passed with a simple additive rule, probably are explicitly multiplying weight times distance.

Although explicit rules are used in some cases, it seems equally clear that they are not used in every case. We know that children and adults will often say one thing and do another, and the verbal reports are incomplete (they do not, for example, encompass small systematic deviations from the rules nor do they accomodate the torque-difference effect). The possibility remains that children (and adults) use implicit rules of the kind that linguists have long believed to underlie the use of natural language.

What connectionist models contribute here is the observation that there is a continuous space of cognitive states, only some of which correspond to what could felicitously be called an implicit rule. Connectionist models can implement rules to any arbitrary desired degree of precision, and when such models are trained in pure environments (i.e., on environments in which the stimuli all embody some system of rules, with no exceptions), they often learn to implement these rules sufficiently precisely that it makes perfectly good sense to describe their behavior in terms of the rules they have learned to

implement. For example, Cleeremans et al. (1991) trained a connnectionist network on strings generated by a simple finite-state grammar and found that the network learned to mimic the predictions of the rules of the grammar. With sufficient training, networks can sometimes converge as closely as desired to the exact predictions of various grammars. Crucially, though, connectionist models can also implement input–output mappings that occupy many other points in a continuous state space of input–output mappings. As a result, such models are capable of making smooth transitions from conforming to one rule to conforming to another.

To summarize the discussion thus far: Implicit knowledge embedded in connectionist networks may correspond very closely to some specified system of rules, but need not, and overt behaivor can be guided either by implicit or explicit knowledge, or perhaps by some weighted combination of the two.

Much more could be said about the exact circumstances under which implicit or explicit knowledge will be used, to what extent these different kinds of knowledge will be blended in different task situations, and to what extent implicit knowledge is profitably describable as capturing some implicit rule. There is another issue, however, that lies closer to the heart of the matter and divides connectionist and nonconnectionist approaches to cognitive development. Explicit knowledge does appear to develop, and in the balance scale arena, Siegler's (1976) data show that it develops in some approximate correspondence with the actual ability to use weight and distance information to predict which side of the scale will go down. Connectionist models may provide a way of accounting for the development of this ability as an implicit skill, but still leave many developmentalists unsatisfied because they say relatively little about why it is that explicit knowledge develops. Why, for example, do 6-year-olds report that they choose the side with the greater weight? Why do 9-year-olds report that they take both weight and distance into account?

It may seem at first sight that the connectionist approach fails to provide any basis for understanding these developments in explicit cognition. However, we have seen that connectionist models of the kind discussed here are actually learning, not just how to predict which side will go down in the balance scale task, but how to represent the relevant dimensions at some internal, cognitive level. Karmiloff-Smith (personal communication, 1992) is exploring the idea that those parts of our cognitive systems that formulate and test explicit rules and generate explicit verbal reports might "see" these representations as inputs. Before the ability to form these representations develops, there would be nothing for explicit cognitive processes to build on, but once the ability to represent some information is learned, it would be available for incorporation into verbal reports and for use in the formulation of explicit rules. Verbal communication would ensure that different individuals observing the same events would describe them in

similar ways, thereby contributing to the development of the ability to translate the represented information into explicit verbal form. In any case, the development of such representations would be a necessary condition for the incorportation of the information they capture into explicit rules, and we would be part of the way toward an understanding of where explicit representations come from.

Obviously, the preceeding paragraph offers only the faintest glimpse of a possible future rapprochement between theories of implicit and explicit cognition. Karmiloff-Smith and Halford are exploring ways to characterize the linkage between implicit and explicit cognitive processes, but this work is in the early stages of development.

CONCLUSION

For the time being, the fact remains that connectionist models have the most to say about implicit rather than explicit forms of cognition. Within this domain, they provide the prospect of allowing us to begin to understand how cognitive change may arise through the gradual cumulating effects of experience. At the same time, their ability to capture many of the main findings from domains that were once thought to lie in the heart of symbolic cognition suggests that the domain of implicit knowledge, and therefore the potential domain of application of connectionist models, may be broad. Surely humans have explicit knowledge and reason with it, but how much of their reasoning is of this explicit form, and how much this explicit form of reasoning depends on underlying implicit knowledge, remains to be seen. Whatever the final resolution of these issues, it seems likely that connectionist models will contribute to their ongoing exploration.

ACKNOWLEDGMENTS

Preparation of this chapter was supported by grants MH–00385 and MH–47566. I thank Robert Siegler and Graeme Halford for useful input.

REFERENCES

Anderson, N. H. (1991). *Contributions to integration theory. Vol. III: Developmental.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Bates, E., & MacWhinney, B. (1987). Competition, variation, and language learning. In B. MacWhinney (Ed.), *Mechanisms of language acquisition* (pp. 157–193). Hillsdale, NJ: Lawrence Erlbaum Associates.

Chauvin, Y. (1989). Toward a connectionist model of symbolic emergence. In *Proceedings of the Eleventh Annual Conference of the Cognitive Science Society* (pp. 580–587). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cleeremans, A., Servan-Schreiber, D., & McClelland, J. (1991). Finite state automata and simple recurrent networks. *Neural Computation, 1,* 372–381.

Elman, J. (1991). Incremental learning, or the importance of starting small. In *Proceedings of the Thirteenth Annual Conference of the Cognitive Science Society* (pp. 443–448). Hillsdale, NJ: Lawrence Erlbaum Associates.

Ferretti, R. P., & Butterfield, E. C. (1986). Are children's rule-assessment classifications invariant across instances of problem types? *Child Development, 57,* 1419–1428.

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

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

Grossberg, S. (1987). Competitive learning: From interactive activation to adaptive resonance. *Cognitive Science, 11,* 23–64.

Halford, G. S. (1993). *Children's understanding: The development of mental models.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Halford, G. S., Wilson, W. H., Guo, J., Gayler, R. W., Wiles, J., & Stewart, J. E. M. (1994). Connectionist implications for processing capacity limitations in analogies. In K. J. Holyoak & J. Barnden (Eds.), *Advances in connectionist and neural computation theory. Vol. 2: Analogical connections* (pp. 363–415). Norwood, NJ: Ablex.

Hinton, G. E. (1981). Implementing semantic networks in parallel hardware. In G. E. Hinton & J. A. Anderson (Eds.), *Parallel models of associative memory* (pp. 161–188). Hillsdale, NJ: Lawrence Erlbaum Associates.

Hinton, G. E. (1989). Deterministic Boltzmann learning performs steepest descent in weight-space. *Neural Computation, I,* 153.

Hinton, G. E., & McClelland, J. L. (1988). Learning representations by recirculation. In D. Z. Anderson (Ed.), *Neural information processing systems* (pp. 358–366). New York: American Institute of Physics.

Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence* (A. Parsons & S. Milgram, Trans.). New York: Basic Books. (Original work published 1955)

Jenkins, E. A. (1989). Knowledge restructuring and cognitive development: A parallel distributed processing approach. In M. A. Luszcz & T. Nettelbeck (Eds.), *Psychological development: Perspectives across the life-span* (pp. 205–216). Amsterdam: North-Holland.

Karmiloff-Smith, A. (1986). From megaprocesses to conscious access: Evidence from children's metalinguistic and repair data. *Cognition, 23,* 95–147.

Karmiloff-Smith, A. (1992). *Beyond modularity: A developmental perspective on cognitive science.* Cambridge, MA: MIT Press.

Klahr, D., & Siegler, R. S. (1978)., The representation of children's knowledge. In H. W. Reese & L. W. Lipsitt (Eds.), *Advances in child development* (Vol. *12,* pp. 61–116). New York: Academic Press.

MacWhinney, B., & Leinbach, J. (1991). Implementations are not conceptualizations:Revising the verb learning model. *Cognition, 40,* 121–157.

MacWhinney, B., Leinbach, J., Taraban, R., & McDonald, J. (1989). Language learning: Cues or rules? *Journal of Memory and Language, 28,* 255–277.

Mareschal, D., & Shultz, T. R. (1993). A connectionist model of the development of seriation. In *Proceedings of the 15th Annual Conference of the Cognitive Science Society* (pp. 676–681). Hillsdale, NJ: Lawrence Erlbaum Associates.

McClelland, J. L. (1989). Parallel distributed processing: Implications for cognition and development. In R. Morris (Ed.), *Parallel distributed processing: Implications for psychology and neurobiology* (pp. 9–45). New York: Oxford University Press.

McClelland, J. L. (1991)., Stochastic interactive processes and the effect of context on perception. *Cognitive Psychology, 23,* 1–44.

McClelland, J. L. (1993). Toward a theoryi of information processing in graded, random, interactive networks. In D. E. Meyer & S. Kornblum (Eds.), *Attention & Performance XIV: Synergies in experimental psychology, artificial intelligence and cognitive neuroscience* (pp. 655–668). Cambridge, MA: MIT Press.

McClelland, J. L., & Jenkins, E., (1991). *Nature, nurture, and connections: Implications of connectionist models for cognitive development.* In K. Van Lehn (Ed.), *Architectures for intelligence* (pp. 41–73). Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Movellan, J. R., & McClelland, J. L. (in press). Learning continuous probability distributions with symmetric diffusion networks. *Cognitive Science.*

Peterson, C. (1991). Mean field theory neural networks for feature recognition, content addressable memory and optimization. *Connection Science, 3,* 3–33.

Pinker, S. (1991). Rules of language. *Science, 253,* 530–535.

Plaut, D., & McClelland, J. (1993). Generalization and componential attractors: Word and nonword reading in an attractor network. In *Proceedings of the 15th Annual Conference of the Cognitive Science Society* (pp. 824–829). Hillsdale, NJ: Lawrence Erlbaum Associates.

Plunkett, K., & Marchman, V. (1989). *Pattern association in a back propagation network: Implications for child language acquisition* (Tech. Rep. No. 8902). San Diego: University of California, Center for Research in Language.

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 2: Current research and theory* (pp. 64–99). New York: Appleton-Century-Crofts.

Rosenblatt, F. (1959). Two theorems of statistical separability in the perceptron. In *Mechanisation of Thought Processes: Proceedings of a Symposium Held at the National Physical Laboratory, November 1958: Vol. 1* (pp. 421–456). London: HM Stationery Office.

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

Rumelhart, D. E., & McClelland, J. L. (1986). PDP models and general issues in cognitive science. In D. E. Rumelhart, J. L. McClelland, & the PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition: Vol. 1* (pp. 110–146). Cambridge, MA: MIT Press.

Rumelhart, D. E., & McClelland, J. L. (1987). Learning the past tenses of English verbs: Implicit rules or parallel distributed processing? In B. MacWhinney (Ed.), *Mechanisms of language acquisition* (pp. 157–193). Hillsdale, NJ: Lawrence Erlbaum Associates.

Rumelhart, D. E., McClelland, J. L., & the PDP Research Group. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition: Vols. 1 & 2.* Cambridge, MA: MIT Press.

Schmidt, W. C., & Shultz, T. R. (1991). An investigation of balance scale success. In *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society* (pp. 72–77). Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Shultz, T. R., Mareschal, D., & Schmidt, W. C. (in press). Modeling cognitive development on balance scale phenomena. *Machine Learning.*

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

Siegler, R. S. (1981). Developmental sequences within and between concepts. *Monographs of the Society for Research in Child Development, 46*(2, Serial No. 189).

Siegler, R. S. (1983). Five generalizations about cognitive development. *American Psychologist, 38,* 263–277.

Siegler, R. S. (in preparation). *Beyond the immaculate transition: Variability, choice, and cognitive development.* New York: Oxford University Press.

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.

Siegler, R. S., & Munakata, Y. (1993, Winter). Beyond the immaculate transition: Advances in the understanding of change. *SRCD Newsletter* pp. 3, 10, 11, 13.

Smolensky, P. (1990). Tensor product variable bginding and the representation of symbolic structyures in connectionist systems. *Artificial Intelligence, 46*(1–2), 159–216.

Spelke, E. S., Breinlinger, K., McComber, J., & Jacobsen, K. (1992). Origins of knowledge. *Psychological Review, 99,* 605–632.

St. John, M. F., & McClelland, J. L. (1987). Reconstructive memory for sentences: A PDP approach. *Proceedings of the Ohio University Inference Conference, 1986* (pp. 270–279). Athens: University of Ohio.

Touretzky, D. S., & Hinton, G. E. (1988). A distributed connectionist production system. *Cognitive Science, 12,* 423–466.

White, H. (in press). Parametric statistical estimation with artificial neural networks. In Y., Chauvin & D. E. Rumelhart (Eds.), *Back-propagation theory, architectures, and applications.* Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Wilkening, F., & Anderson, N. H. (1982). Comparison of two rule-assessment methodologies for studying cognitive development and knowledge structure. *Psychological Bulletin, 92,* 215–237.

Wilkening, F., & Anderson, N. H. (1991). Representation and diagnosis of knowledge structures in developmental psychology. In N. H. Anderson (Ed.), *Contributions to integration theory. Vol. 3: Developmental* (pp. 45–80). Hillsdale, NJ: Lawrence Erlbaum Associates.