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

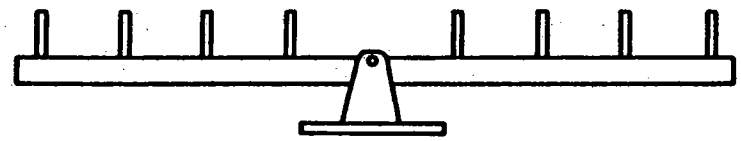

FIG. 4.1. A balance scale of the type used by Siegler (1976, 1981). Reprinted from Figure 1 of Siegler (1976), with permission.

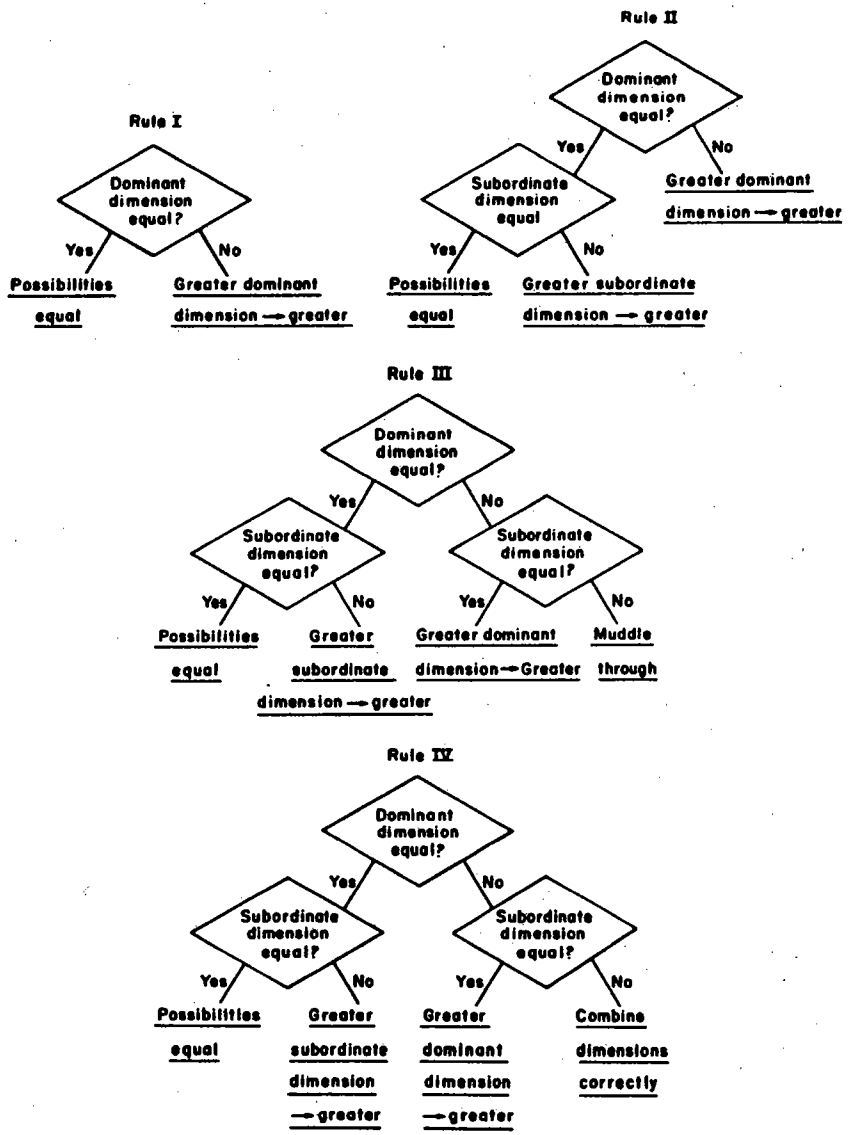

FIG. 4.2. The four Rules identified by Siegler (1976). Reprinted from Figure 1 of Siegler (1981), with permission.

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.

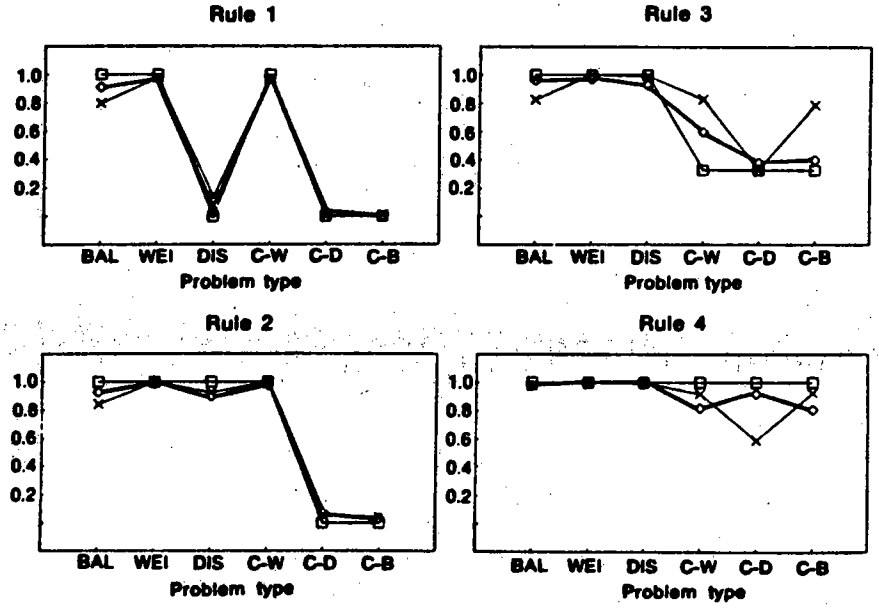

FIG. 4.4. Patterns of performance on each of the six problem types by children (circles), and the McClelland (1989) balance scale model (Xs) averaged over children or networks satisfying Siegler's (1981) criteria for use of each rule. Also shown is the pattern of performance corresponding to exact adherence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

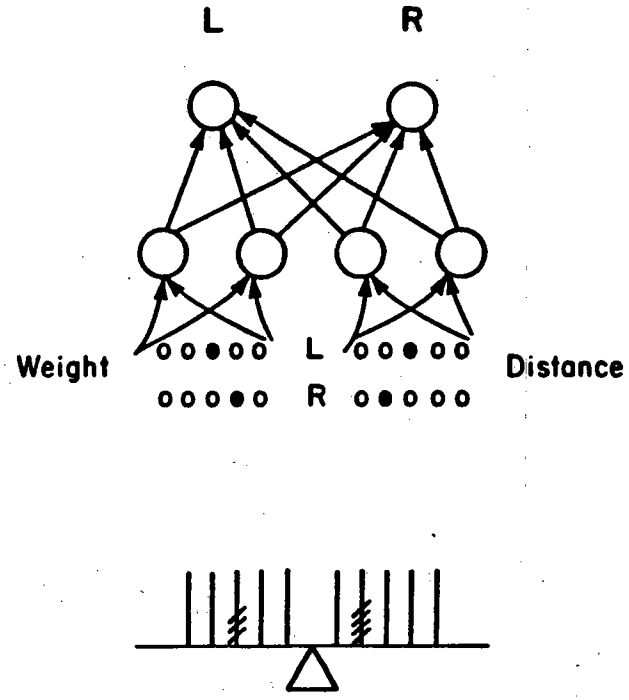

FIG. 4.5. Connectionist network used in the McClelland (1989) balance scale model. Reprinted from Figure 2.7 of McClelland (1989), with permission.

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