
Modeling Cognitive Development With a Generative Connectionist Algorithm

Thomas R. Shultz
William C. Schmidt
David Buckingham
Denis Mareschal
McGill University

One of the key unsolved problems in cognitive development is the precise specification of developmental transition mechanisms. As the work in this volume attests, it is clear that computational modeling can provide insights into this problem. In this chapter, we focus on the applicability of a specific generative connectionist algorithm, cascade-correlation (Fahlman & Lebiere, 1990), as a process model of transition mechanisms. Generative connectionist algorithms build their own network topologies as they learn, allowing them to simulate both qualitative and quantitative developmental changes. We compare and contrast cascade-correlation, Piaget's notions of assimilation and accommodation, Papert's little known but historically relevant *genetron* model, conventional back-propagation networks, and rule-based models.

Specific cascade-correlation models of a wide range of developmental phenomena are presented. These include the balance scale task; concepts of potency and resistance in causal reasoning; seriation; integration of the concepts of distance, time, and velocity; and personal pronouns. Descriptions of these simulations stress the degree to which the models capture the essential known psychological phenomena, generate new testable predictions, and provide explanatory insights. In several cases, the simulation results underscore clear advantages of connectionist modeling techniques. Abstraction across the various models yields a set of domain-general constraints for cognitive development. Particular domain-specific constraints are identified. Finally, the models demonstrate that connectionist approaches can be successful even on relatively high-level cognitive tasks.

TRANSITION AS A MAJOR UNSOLVED PROBLEM

Although researchers have come some distance in understanding the development of children's thinking, much of the research has been directed toward a structural analysis of the relevant thought processes. Mechanisms governing developmental transitions were often neglected as this issue was typically viewed as too complex to fathom. Nevertheless, it has also been argued that such structural descriptions do not suffice (Boden, 1982). To fully understand cognitive development, one also needs a theory of the various transformations that structures undergo. Moreover, even when transition mechanisms were tentatively proposed, they tended to account for only qualitative transitions (van Geert, 1991).

Piaget's theory of cognitive development is a prime example of a theory that fails to specify transition mechanisms with sufficient precision (Bates & Elman, 1993; Boden, 1982). The proposed motors of development in Piagetian theory are assimilation, accommodation, and equilibration (Piaget, 1972). Assimilation consists in the child modifying the incoming environmental information to allow it to fit within the child's existing structures. Accommodation occurs when the child modifies existing mental structures under environmental pressures. Finally, equilibration consists in the coordination of assimilation and accommodation so as to achieve optimal harmony between the environment and mental structures. Assimilation and accommodation always occur together, although phases of predominant assimilation or predominant accommodation also occur.

Piaget struggled throughout his career to precisely formulate these ideas (Piaget, 1977). Yet many contemporary researchers feel that Piaget fell short of this goal (Boden, 1982). Piaget was criticized for assuming what he was trying to explain (Macnamara, 1976), as well as for not analyzing critical presuppositions (Ninio, 1979). The issue of how newly created structures are integrated with older structures, without completely disrupting the child's existing reasoning abilities, is not resolved (Boden, 1982). Piaget's concepts are not constrained enough to carry over into the computational domain.

Although we have focused on Piaget's work to illustrate our point, he is not the only theorist who faltered when faced with the problem of transition mechanisms. Indeed, when reviewing a number of proposed transition mechanisms, Flavell (1984) found fault with all of them. The criteria he suggested for evaluating theories of transition are as follows. First and foremost, a good theory should propose clearly defined mechanisms. All modes of operation should be described precisely and in great detail. Second, the theories should suggest empirical studies that would allow evaluation of the plausibility of the models as accounts of how cognitive development proceeds. Thus, the ultimate value of a theory is related to its

ability to account for existing observations and to suggest new studies that lead to the discovery of novel phenomena.

A COMPUTATIONAL APPROACH IS THE REQUIRED SOLUTION

One solution to the lack of precision in transition mechanisms is to use a formal language such as mathematics to describe the processes involved. The consistency of the proposed mechanisms could then be tested through computer simulations. Simulation is an ideal medium for exploring the implications of a complex model and can result in the prediction of seemingly counterintuitive findings (Lewandowsky, 1993). It provides a formal framework that can disambiguate verbal formulations.

The connectionist models we present here make use of both the mathematical and computational levels. The dynamics of a network are specified by activation functions, learning rules, training regimes, and so on. The legitimacy of the proposed model is then evaluated through explicit comparisons of computer simulations with observed psychological data.

The complementarity of computational modeling and traditional developmental theorizing promises to be fruitful. If modelers can take account of the empirical data provided by traditional psychological accounts, then crucial questions may be answerable (Boden, 1980). The promise of the computational approach is that it naturally satisfies Flavell's (1984) methodological criteria because it is precise and generates novel predictions.

With these goals in mind, early modelers of cognitive development often adopted a production system approach (Boden, 1988). Initially these rule-based models were static descriptions of performance during a particular stage of development. More recently though, developmental researchers are exploring the utility of self-modifying production systems (Klahr, Langley, & Neches, 1987; Newell, 1990). The adequacy of these models, however, was challenged by the application of connectionist models.

Attempts to provide a formal analysis and implementation of Piagetian development within a connectionist framework can be traced to the early 1960s. Papert (1963) tried to build an automated equilibratory system that he called the genetron. He argued that behavioristic learning theorists were correct in claiming that the underlying mechanisms of intelligence are simple if considered independently, but that they unfortunately ignored the interactive complexity of the ensemble. Papert argued that models that relied on progressive and hierarchical acquisition of functions would evolve in a stagelike manner. He hypothesized that such models would develop through alternating periods of first lowering variability in the system, and

then constructing new functions. The genetron was largely hand constructed from a collection of elementary perceptrons (two-layered networks). Although an attempt was made to simulate children's developing integration of information in a length comparison task, no tangible empirical results were provided and the project was apparently dropped. Presumably, the technical limitations of perceptrons (Minsky & Papert, 1969) outweighed the potential benefits of this type of model.

APPROPRIATENESS OF CONNECTIONIST APPROACHES

Modern connectionist methods offer a number of potential advantages for the creation of process models of cognition, including the ability of these nets to learn procedural and declarative knowledge, generalize to novel situations, and derive coherent solutions despite variable environmental input. Because of these strengths, connectionist simulations are now starting to illuminate several aspects of cognitive, perceptual, and language development.

Harnad, Hanson, and Lubin (1994) showed how categorical perception might arise as a natural side effect of back-propagation learning. Several psychological phenomena in concept acquisition and semantic development were addressed within the connectionist framework, including prototype and typicality effects, semantic over- and underextension, the mutual exclusivity constraint, vocabulary explosion, and the emergence of comprehension before production (Chauvin, 1989; Schyns, 1991).

Several other aspects of language development were also simulated with connectionist techniques, including the formation of the English past tense (Hare & Elman, 1992; Marchman, 1992; Plunkett & Marchman, 1991), article choice for German nouns (MacWhinney, Leinbach, Taraban, & McDonald, 1989), word recognition and naming (Seidenberg & McClelland, 1989), and syntactic development (Elman, 1993).

The pioneering attempt to apply modern connectionist techniques to developmental problem solving tasks was McClelland's (1989) model of balance scale stages, a task that we discuss in detail later.

Taken together, this research suggests that a connectionist approach to cognitive development cannot be easily dismissed. The models yielded qualitatively accurate simulations of a variety of different phenomena, and provided a number of explanatory insights. Recent theoretical papers argued that the application of connectionist techniques to cognitive development is fostering a needed return to the traditional issues of change and transition (Bates & Elman, 1993; Plunkett & Sinha, 1992; Shultz, 1991). For the aforementioned reasons of difficulty, developmental psychologists have tended to ignore issues of change and transition in favor of diagnostic

concerns. Connectionist networks provide a precise and concrete way to think about developmental change.

All of the foregoing models had static network processing structures that were hand designed by the researchers. Further, all of the learning was accomplished solely by small adjustments in network weights over many epochs. For our simulations of cognitive development, we opted instead for a generative algorithm, in which small quantitative changes in connection weights are augmented by qualitative changes in network topology as learning progresses. A number of generative connectionist learning frameworks exist (Alpaydin, 1991; Hertz, Krogh, & Palmer, 1991). We focus on one, cascade-correlation (Fahlman & Lebiere, 1990), that is particularly suitable for modeling cognitive development.

Continuous small weight changes can sometimes produce qualitative behavioral shifts, even in static networks. Such outcomes were described in terms of mathematical catastrophe theory (Pollack, 1990; van der Maas & Molenaar, 1992). A traditional view of cognitive development is that qualitative behavioral changes arise instead from a major restructuring of cognitive processing (Piaget & Inhelder, 1969). A possible advantage of using generative network models is that both types of transition mechanisms can be examined simultaneously. Some qualitative behavioral changes may result from continuous quantitative adjustments alone, whereas others may also require qualitative changes in network topology. Research with static networks does not facilitate the study of interactions between underlying quantitative and qualitative changes.

THE CASCADE-CORRELATION LEARNING ALGORITHM

Cascade-correlation begins with a minimal network topology consisting of a single layer of input units fully connected to a single layer of output units (Fahlman & Lebiere, 1990; Fig. 5.1a). The algorithm then designs and recruits its own hidden units as and when it needs them. Cascade-correlation operates in two alternating phases: an output phase in which weights leading to output units are modified (Figs. 5.1a, 5.1c) and an input phase in which candidate hidden units are trained for installation in the network (Fig. 5.1b). The name *cascade-correlation* presumably derives from the way that hidden units are recruited into the network. In each input phase, the candidate hidden unit whose activations correlate maximally with the network's existing error is selected for installation. Hidden units are arranged in a sort of cascade, so that each new hidden unit receives input from all of the previous hidden units.

Learning proceeds in batch mode, that is, all weight modifications occur

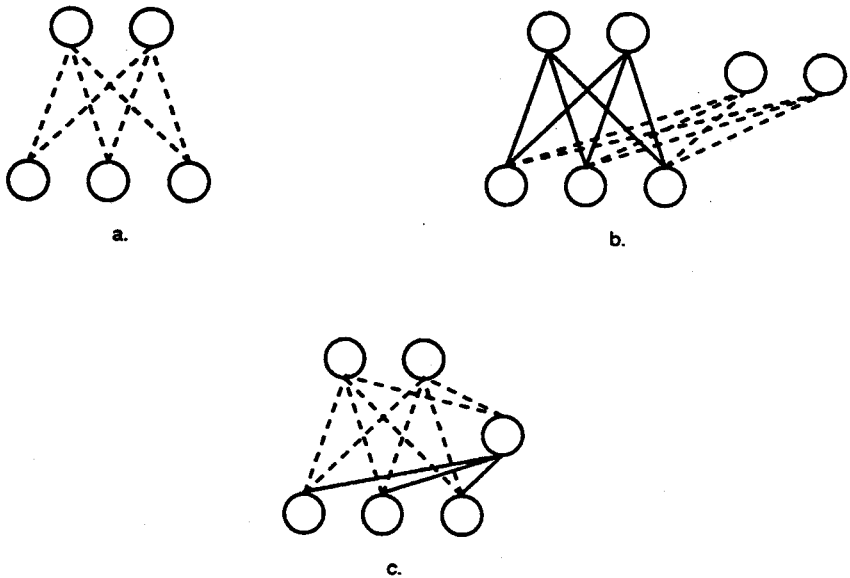


FIG. 5.1. Training in cascade-correlation. Modifiable connections are represented by dashed lines and nonmodifiable connections are represented by solid lines. (1a) and (1c) refer to output phases, (1b) to an input phase. Adapted with permission from Shultz, Mareschal, and Schmidt (1994).

after a complete blocked presentation of all of the input-output pattern pairs. This is a requirement of the quickprop weight adjustment algorithm that is used within cascade-correlation (Fahlman, 1988). Such a complete presentation of the training patterns is called an *epoch*. *Victory* is achieved when all output activations are within a small threshold of their target values.

There is considerable psychological (Oden, 1987) and physiological (Dudai, 1989; Squire, 1987) evidence for batch learning. For example, the hippocampus processes information in batch mode in order to relay its information to relevant cortical areas at some later time. Batch learning is potentially more computationally efficient than pattern learning because it requires fewer weight updates for the same number of patterns.¹ Batch learning also avoids making and unmaking redundant weight changes that might result from the purely local evaluations of error signals in pattern learning. Even in batch learning, however, outputs are compared to their targets independently of other patterns. Thus, the system never has to process more than one pattern at a time although it does keep a running sum of the error that is eventually used to adjust the weights.

¹In pattern learning, weights are adjusted after every training pattern.

Output Training

During the output training phase, weights leading to the output units are modified so as to minimize the sum of squared error (E):

$$E = \sum_o \sum_p (A_{op} - T_{op})^2 \quad (1)$$

where o indexes the output units, p indexes the input-output pairs, A is the actual activation of an output unit, and T is the target activation for that output unit. If E stagnates (i.e., ceases to change by more than a specified amount for a certain number of epochs) or a specified maximum number of epochs elapses, the algorithm changes to the input phase.

Input Training

During the input training phase, weights leading to the output units are frozen, meaning that they are no longer allowed to change. A number of candidate hidden units are connected with random weights from all input units and existing hidden units. The weights leading to each candidate unit are then adjusted to maximize the absolute value of the correlation (C) between the activation of that unit and the residual error at the output units, across all patterns.

$$C = \frac{\sum_o \sum_p |(h_p - \langle h \rangle)(e_{op} - \langle e_o \rangle)|}{\sum_o \sum_p (e_{op} - \langle e_o \rangle)^2} \quad (2)$$

where h_p is the activation of the candidate hidden unit for pattern p , $\langle h \rangle$ is the mean activation of the candidate hidden unit for all patterns, e_{op} is the residual error at output o for pattern p , and $\langle e_o \rangle$ is the mean residual error at output o for all the patterns.

Training continues until C stagnates or a prespecified maximum number of epochs has elapsed as described earlier. At this point, the candidate unit with the largest C is retained and all other candidate units are discarded. The input weights to the newly installed hidden units are then frozen and the unit is allowed to send output to all of the output units. The algorithm returns to the output training phase with the added power of the new hidden unit that is particularly adept at detecting the residual error that the network was encountering.

The Quickprop Algorithm

Rather than using the usual back-propagation algorithm (Rumelhart, Hinton, & Williams, 1986) to modify weights, cascade-correlation uses a second-order algorithm developed by Fahlman (1988) called *quickprop*. Quickprop was inspired by Newton's minimization methods and incorporates curvature information into the optimization process. It is based on two assumptions: that weights contribute independently to the function being optimized, and that the function is locally quadratic. The value of the function and its slope at the current and previous points are used to uniquely define a parabola. The weight that minimizes this parabola is then selected as the next weight for that connection in the network. Although this is the mechanism at the heart of the quickprop algorithm, under some conditions modifications occur so as to bootstrap the process and avoid certain computational pitfalls (for a detailed description and justification of these modifications see Mareschal, 1992). The actual update rules in Fahlman's code are:

$$\begin{aligned}
 w_3 - w_2 &= \epsilon f(w_2) \text{ if } w_2 - w_1 = 0 \\
 w_3 - w_2 &= \frac{f(w_2)}{f(w_1) - f(w_2)} (w_2 - w_1) \text{ if } w_2 - w_1 \neq 0 \text{ and} \\
 &\quad \left| \frac{f(w_1)}{f(w_1) - f(w_2)} \right| < \mu \\
 w_3 - w_2 &= \mu (w_2 - w_1) \text{ otherwise}
 \end{aligned} \tag{3}$$

where the indices 1, 2, 3, represent three consecutive time steps, f is the derivative of the function being optimized (E in the case of the output phase, C for the input phase), ϵ is a parameter controlling the amount of gradient descent, and μ is a parameter controlling the maximum step size. The product $\epsilon f(w_2)$ is added to the weight update even when the previous weight change is nonzero, that is, in lines 2 and 3 of Equation 3, except when the current slope is of opposite sign from the previous slope. This detail is not presented in Equation 3 in order to keep this equation legible.

Activation Functions

Three types of activation functions are available for hidden units in cascade-correlation: *linear*, *sigmoid*, and *gaussian*. Throughout all of the present models, we used sigmoid activation functions, symmetrical around 0 and ranging from -0.5 to +0.5.

$$y_i = \frac{1}{1 + \exp\left(-\sum_j w_{ij}x_j\right)} - 0.5 \quad (4)$$

where y is the resulting activation of the receiving unit indexed by i , x is the activation of a sending unit indexed by j , and w is the weight connecting those two units. Our input units typically have linear activation functions, meaning that they sum all input into them and output that sum.

Developmental Implications

Papert's (1963) genetron, developed specifically to model Piagetian phenomena, is a historical precedent for cascade-correlation. The genetron consisted of hierarchically ordered and recurrently connected perceptron units. Papert gave several mathematical arguments for why this model should develop through alternate phases of noise reduction and internal function construction, thereby giving rise to stagelike development.

Mareschal (1991) identified similarities between the genetron and cascade-correlation. Both can be expressed within the Piagetian framework. For Piaget, equilibration consisted of alternating periods of accommodation to new information followed by assimilation of familiar information. In connectionist models, the knowledge structure of a domain is embodied in the nodal architecture and the knowledge content is embodied in the weights linking those nodes. The period of error reduction can be viewed as the assimilation (or partial assimilation) of information into previously existing knowledge structures. Only the weights (i.e., the content of the knowledge) are being modified. The period of hidden unit recruitment can be seen as the accommodation of knowledge structures to unassimilated information. As with the child, assimilation corresponds to a period in which new information can be integrated within existing knowledge structures, whereas accommodation corresponds to a period in which genuinely new structures are created out of older ones without functional impairment of the system as a whole.²

We now turn to a review of some of our cascade-correlation models of developmental phenomena. These include models of balance scale phenomena; concepts of potency and resistance in causal reasoning; seriation; integration of the dimensions of distance, time, and velocity; and acquisition of personal pronouns. All of these simulations, except for pronouns, involve reasoning about aspects of the physical world. As noted earlier,

²We extend this interpretation of cascade-correlation in terms of assimilation and accommodation in the General Conclusions and Discussion section.

most developmental connectionist work concerned concept or language acquisition. The simulation choices reflect a smattering of a benchmark modeling problem (balance scale), topics we worked on and thus were interested in and knew well (potency and resistance; pronouns), and basic well-known developmental phenomena (seriation; distance, time, velocity).

THE BALANCE SCALE

The balance scale task is an appealing candidate for cognitive developmental computational modeling. The task combines an explicit and well-defined methodology with detailed human observations. The clarity and replicability of balance scale phenomena with humans, coupled with the classical developmental appeal of its stagelike character, led to both connectionist (McClelland, 1989; Shultz & Schmidt, 1991) and rule-based models (Langley, 1987; Newell, 1990).

We used cascade-correlation as a transition mechanism to create two general types of working models of developmental balance scale phenomena, each embodying different sets of theoretical assumptions. One type of model adopted the assumption of a biased training environment, after McClelland (1989). A second type of model investigated the effect that prestructuring the network's starting state had on development of behavior on the balance scale. After reviewing the task's psychological background, an evaluation of existing balance scale models is presented, and then we report on the two cascade-correlation models.

Psychology of the Balance Scale

The balance scale task was developed by Inhelder and Piaget (1958) for their studies of proportionality concepts. Examples of the balance scale apparatus appear on the left side of Fig. 5.2. The child is shown a balance scale supported by blocks so that the scale stays in the balanced position. Next, a number of weights are placed around one of a number of evenly spaced pegs on either side of the fulcrum, and it becomes the child's task to predict which arm will go down, or whether the scale will balance, once supporting blocks are removed. For perfect responding, the task requires that the child integrate information from the two dimensions of weight and distance. Perfect performance on this task can be calculated via multiplication. Torques can be calculated for both the left and right arms by multiplying weight by distance; the side with the larger torque will go down. If the torques are equal, then the scale will balance.

Siegler (1976, 1981) operationalized Inhelder and Piaget's (1958) observations with a rule assessment methodology, proposing that development

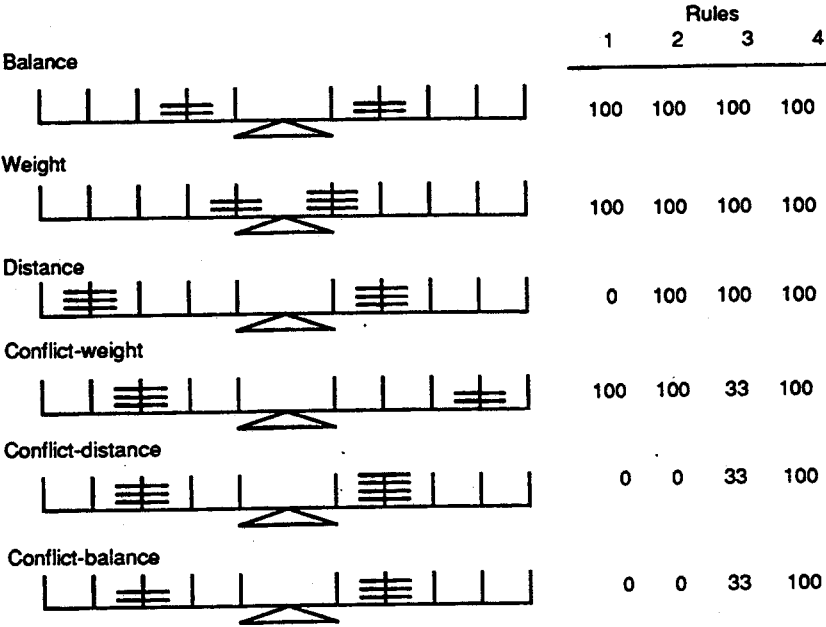


FIG. 5.2. Predicted percentage of correct answers on different balance scale problem types for children responding in accordance with different rules. Adapted with permission from Shultz, Mareschal, and Schmidt (1994).

on the balance scale task is characterized by the use of four increasingly powerful rules. Children diagnosed as using a given rule are classified as being in the corresponding stage of development. Stage 1 performers use only weight information to determine if the scale will balance. Stage 2 subjects emphasize weight information but consider distance if weights on either side of the fulcrum are equal. Stage 3 subjects correctly integrate both weight and distance information for simple problems, but respond indecisively when one arm has greater weight and the other greater distance. Stage 4 subjects correctly integrate weight and distance information for near perfect performance, suggesting but not requiring that they explicitly understand and use torques. There is some debate concerning the proportion of the population that reaches Stage 4, yet it is clear that some individuals do so (Siegler, 1981). Because very few studies have assessed adult competencies, there remains the possibility that Stage 4 performance may be achieved given a high degree of experience, perhaps even without explicit knowledge of torque (Shultz, Mareschal, & Schmidt, 1994).

In order to assess children's stages of development, Siegler partitioned the entire set of balance scale problems into six different problem types, and used performance on a subset of these problems to classify subjects as conforming to a particular rule. *Balance* problems have equal numbers of weights placed at equal distances from the fulcrum. In *weight* problems,

distances on either side of the fulcrum are equal so the side with more weights goes down. In *distance* problems, the arm with greater distance goes down because the two sides have equal weights. *Conflict* problems have greater weight on one arm and greater distance on the other. The correct response to a conflict problem determines its classification as a *conflict-weight*, *conflict-distance*, or *conflict-balance* problem. A child's performance was classified by the pattern of successes and errors observed when tested with 24 problems, 4 from each of the 6 problem types. Siegler's four rules, as they appear on the right side of Fig. 5.2, define the expected percentages of correct responses across the six problem types.

A number of basic observations emerged from balance scale research using both Piagetian style observation and Siegler's rule assessment methodology. First, as children get older, they appear to progress systematically through rule-based stages as described earlier. A second developmental observation is a pattern of U-shaped performance between Stages 2 and 4 for conflict-weight problems. Children predict correct outcomes for these problems during Stages 1 and 2, lose this ability during Stage 3, and regain it in Stage 4. A third balance scale regularity is that the greater the difference in torque between the two sides of a balance scale, the more likely it is that a child will respond correctly (Ferretti & Butterfield, 1986). This torque difference effect makes it possible for the same child to be classified at different stages by Siegler's rule assessment procedure depending on the test problems' differences in torque.

A Review of Previous Balance Scale Models

There are three symbolic models and two connectionist models of the balance scale task published to date. Klahr and Siegler (1978) modeled each of the four stages of balance scale development as production rules. This work is descriptive of the child's performance, but no attempt was made to provide a mechanism for stage transitions. Langley (1987) expanded upon Klahr and Siegler's (1978) findings by adding a transition mechanism. Langley's model started with a set of rules that made random predictions. The system would then learn from its errors on specific problems and improve with experience. The transition mechanism was a discrimination process that looked for differences between cases in which correct predictions were made and cases in which errors were made.

Langley (1987) reported that the system learned to perform at Stage 3, but never reached Stage 4. Moreover, it did not appear to move through Stages 1 and 2 on its way to Stage 3. The model's responses were not tested for the torque difference effect or for U-shaped development on conflict-weight problems.

Newell (1990) reported on a model of the balance scale task using the

Soar architecture, which creates rules by chunking the results of search-based problem solving. This model learned to correctly respond to just four balance scale problems, moving through Stages 1, 2, and 3 in the process, although Stage 4 was never achieved. The model's responses were not tested using Siegler's rule assessment methodology, and no attempt was made to test for the torque difference effect or for U-shaped development on conflict-weight problems. There was no comparison of the model's output to human data, except for an overall qualitative judgment of stage transition. The psychological realism of this Soar model is questionable. First, stage transition apparently occurred from processing a single exemplar, whereas children have years of experience lifting and holding objects before they make the same transition. Second, it is unclear how dependent the Soar model was on observing specific exemplars in a certain order (Shultz & Schmidt, 1991). Finally, as mentioned, the Soar model failed to reach the level of Stage 4 responding.

McClelland (1989) used a back-propagation network that assumed separate processing of weight and distance information, implemented by having two hidden units receive only weight input and two others receive only distance input. A subset of all the possible training patterns was randomly selected each epoch with a strong bias in favor of equal distance problems. It was suggested that this bias, responsible for the appearance of the first two stages, reflects children's extensive experiences picking up differing numbers of objects and their limited experience at placing such objects at various distances from a fulcrum. Recent extensions to this model (McClelland & Jenkins, 1991) also demonstrated a differential readiness to learn, behaviorally similar to children in Siegler's (1976) experiments.

McClelland's model successfully captured many of the details found in the human balance scale literature, including orderly stage progression. It can also capture the torque difference effect (Schmidt & Shultz, 1991). However, this model failed to achieve a consistent level of Stage 4 performance. Of notable merit, McClelland's was the first balance scale model subjected to the rigorous rule assessment methodology used with humans and the first to demonstrate that progression through rule-based stages could be accomplished by connectionist learning.

The Environmental Bias Model

We have previously reported on a cascade-correlation model of development on a 10-peg, 10-weight balance scale task (Shultz, Mareschal, & Schmidt, 1994; Shultz & Schmidt, 1991). This model incorporated the assumption of an environmental training bias (after McClelland, 1989), where equal-distance problems (problems in which the distance of weights on either side of the balance scale are equal) were much more frequent than

other types of problems in the training corpus. This training bias makes it difficult for the network to extract information contributed by the distance dimension, but allows the network to rely on weight information, thus encouraging initial performance at Stages 1 and 2. Coupled with this environmental bias was a training method that gradually introduced new patterns for the network to learn from. This training method is called *expansion training* and it conforms to the authors' assumption that the child's environment changes gradually as the child is exposed to more and more instances of lifting and holding objects.

The initial network topology of the environmental bias model appears in Fig. 5.3. The input encoding of both distance and weight information was implemented using integers in the range of 1 to 5. The activation values of the outputs (2 real numbers between -0.5 and $+0.5$) are interpreted to transform the network's output into one of three possible predictions. A prediction of *left side down* was conveyed by excitation of the first output unit and inhibition of the second output, whereas a prediction of *right side down* was conveyed by the reverse pattern. A *balance* prediction was conveyed by neutral (i.e., 0) values on both outputs.

The environmental bias model's initial training corpus was composed of 100 training patterns randomly selected without replacement from the entire set of 625 possible training instances (1 to 5 weights per peg, crossed with 1 to 5 regularly spaced pegs on either arm). Of these initial patterns, approximately 90% had weights placed equally distant from the fulcrum. On each subsequent epoch, another training pattern was randomly drawn

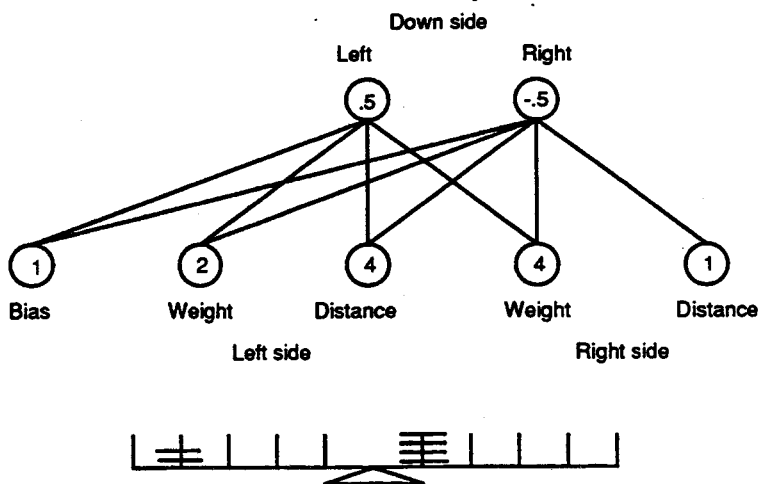


FIG. 5.3. Initial balance scale network topology and an example of input encoding for the environmental bias balance scale network.

with replacement (also subject to the same bias) and added to the set of instances from which the network learned.

After each output epoch, the network's performance was evaluated using Siegler's rule assessment methodology. Twenty-four testing instances balanced for torque difference (four from each of the six different problem types, and of those four, one from each of four different torque difference levels) were randomly chosen at the beginning of each simulation run for use in assessing the network. All 16 runs of the model demonstrated orderly longitudinal stage progression as scored according to the criteria used in human studies (Siegler, 1976). Transitions between stages were typically soft, with a good deal of going back and forth between stages before settling into the higher stage. There were also observations of stage skipping and regression back to earlier stages (Shultz, Mareschal, & Schmidt, 1994; Shultz & Schmidt, 1991). A limited amount of stage skipping and regression is characteristic of human data as well (Chletsos, De Lisi, Turner, & McGillicuddy-De Lisi, 1989; Siegler, 1981).

Also congruent with human data, less error was observed in network responses to testing problems with larger torque differences. Recall that torque difference is the absolute difference between the torques on each arm of the balance scale.

The environmental bias model was also capable of strong Stage 4 performance, a quality missing in all previous balance scale modeling efforts (Langley, 1987; McClelland, 1989; Newell, 1990). Furthermore, the model captured all of these phenomena without having to separate the internal representations of weight and distance, as were required in previous connectionist models of the task (McClelland, 1989).

The Prestructured Weight Dimension Model

A second modeling attempt investigated whether a cascade-correlation model would naturally pass through all of the stages witnessed with humans if it were to start off focusing on weight information. This simulation tested the merits of a nativist position that sees evolution as having innately specified some initial state, as well as the initial structure, of the computational apparatus. This is similar in spirit to Spelke's (1990) suggestion that infants have implicit assumptions about their world that organize their sensory inputs.

Just as the evolutionary medium consists of an infinitely large space of possible tokens, so too does the connectionist medium. The entire space of possible models for a given network topology is delimited by a separate dimension for each degree of freedom in the model (each network connection, or weight), and the starting point of the model within this connection

space constrains the form that the token model can take on the basis of new experiences. Evolution can be viewed in this light, as a coarse search through a space of possible organism types. Learning can be described as the further investigation of the local connection space around a token individual (Nolfi, Elman, & Parisi, 1990).

The current simulation investigated whether or not a connectionist model possessing, from the outset, a structure for assimilating weight but not distance information would produce the developmental sequence observed in children's balance scale performance. Another way of phrasing this is to ask whether domain-specific knowledge, as part of the network's starting state, can produce a realistic developmental sequence as domain-general learning procedures are applied. Whereas the environmental model discussed earlier places domain-specific constraints in the environment, this model internalizes such constraints.

There are several ways of placing a network into a particular region of connection space. One could supply initial connection weight values by hand, arrange for them to be inherited by natural selection (Belew, McNery, & Schraudolph, 1990), or pretrain the network with equal distance problems.

We adopted the last of these approaches for ease of implementation, creating a model in which networks were first placed in a region of connection space such that they performed at the level of stage 1, before being exposed to a corpus of unbiased training instances. The initial prestructured network topology appears in Fig. 5.4. Ten inputs and an obligatory bias unit were fully connected to each of two output units. Of the 10 inputs, 5 represented the left arm of the balance scale, and 5 represented

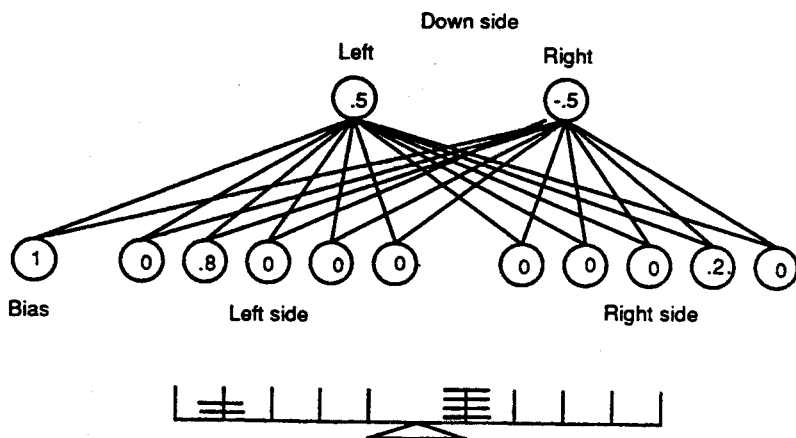


FIG. 5.4. Initial network topology and an example of input encoding for the prestructured weight balance scale network.

the right arm. Distance for either arm was encoded by a real valued number in the range of 0.0 to 1.0 that was proportional to the distance at which weights were placed on the arm. The weight dimension was represented locally with each unit of a given arm corresponding to the number of weights placed on that arm. For a given balance scale problem, the distance input value was entered as input to the corresponding weight input unit. This input encoding provides a compact representation of balance scale problems, yet does not produce a significantly different style of learning from locally encoded networks (Schmidt, 1991). The method of interpreting the outputs used in the environmental bias model was retained for use in the current model.

Network weights were randomly initialized in the range of -1.0 to $+1.0$ before pretraining began. During pretraining, the networks were exposed to a corpus consisting only of weight problems, until a proficient level of performance was reached, thereby endowing the network with a structure capable of processing weight problems. At this point, the network's connection weights were preserved and the training set was switched to include all 625 possible training instances. This second phase of training continued until the network had learned all of the training patterns.

After each output epoch during the second training phase, the responses of 96 networks to a set of 88 different testing problems were recorded. The testing problems included the 24 testing problems used by McClelland (1989) to diagnose network performance, and 4 problems from each of the four nonbalance problem types at four different torque difference levels. This collection of testing problems provided enough information to classify the networks' responses at four different levels of torque difference, as well as in the conventional manner in which problem type is confounded with torque-difference level.

Longitudinally, of the 96 networks, 83 (86%) were classified by one of Siegler's four rules at some point in development. Eighty nets (83%) demonstrated all four rules in the idealized sequence (1, 2, 3, 4). Sixty-eight nets (71%) displayed temporary regressions from Rule 4 to Rule 3. No other substantial regressions occurred. Thirteen nets (14%) skipped a stage at some point in development.

All of the networks performed well on the weight problems without requiring hidden units, and each required the addition of a single hidden unit in order to accommodate distance information. In all cases, the addition of this unit propelled them from Stage 1 into subsequent stages. This transition demonstrates that at least some qualitative changes in behavior (the use of distance information) require qualitative changes in network structure. However, we also observe qualitative behavioral changes (conforming to increasingly sophisticated rules) from the less drastic quantitative adjustment of connection weights.

Next, a cross-sectional analysis of rule use was carried out according to the method detailed by Schmidt (1991; Schmidt & Shultz, 1991). The 86% of networks that were classified by one of Siegler's four rules was fairly close to the human figure of 78%. If one drops the youngest age group from this calculation, as Siegler (1981) did, then 84% of nets were classified, compared with 91% of children. Fig. 5.5 plots the percentage of errors made on Siegler's six problem types at each of the four stages for Siegler's rules, children's data, and our network simulations. The epochs chosen for network results were those that most closely matched the children's data. In Fig. 5.5 there is a close correspondence between human and network responses for Rules 1 and 2. Rule 3 network performance deviated from the children's data in a number of ways, with poorer performance on weight and distance problems and better performance on conflict-balance prob-

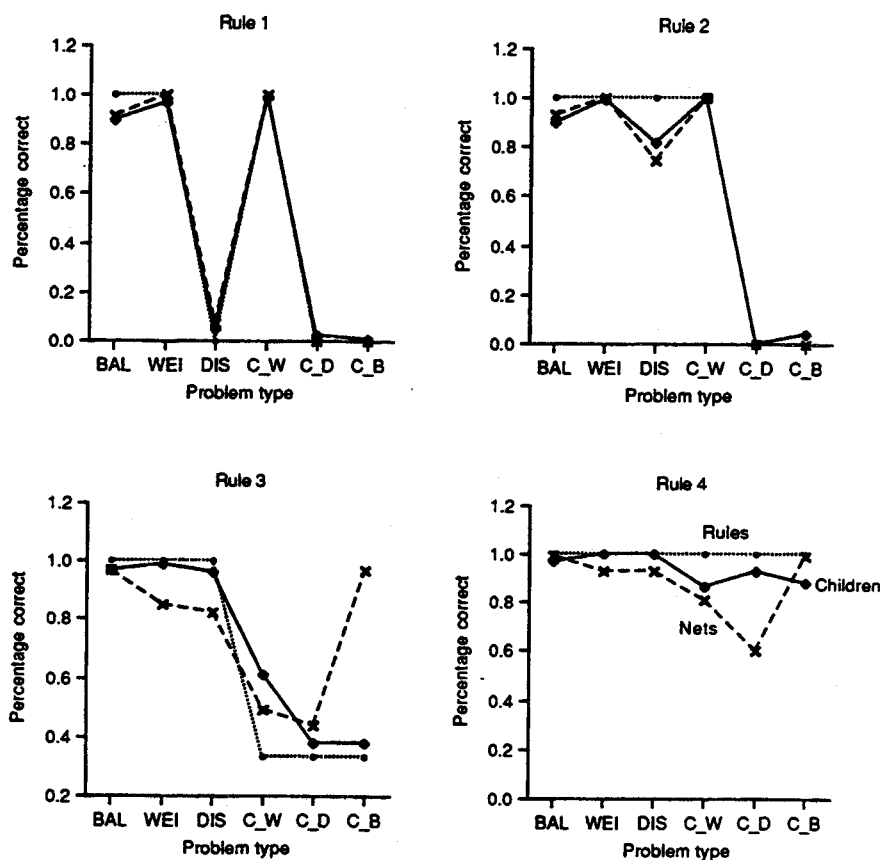


FIG. 5.5. Proportion correct on different balance scale problem types for rules, children, and networks with early weight experience.

lems. Network Rule 4 performance was also generally worse than that of the children, with a notable difference on conflict-distance problems. Fig. 5.5 also depicts a U-shaped developmental trend in network performance on conflict-weight problems across Stages 2 through 4, similar to that observed with children.

Two different analyses for detecting the torque difference effect were performed on these networks. First, a cross-sectional analysis classified the network with four different testing sets, each containing problems of differing magnitudes of torque difference. Classifications of network performance improved with torque difference. As depicted in Fig. 5.6, the model's mean number of problems correct for each of the problem types increased with torque difference, except at torque difference level 4 for weight problems. The human data show similar trends, with children more frequently solving problems correctly from larger torque difference levels (Ferretti & Butterfield, 1986).

The second analysis used to assess the torque difference effect contrasted a network's total sum of squared error scores (*tss*) collected at each level of torque difference midway (output epoch 40) and late (output epoch 80) in training. Networks were tested with each of the four distinct testing sets,

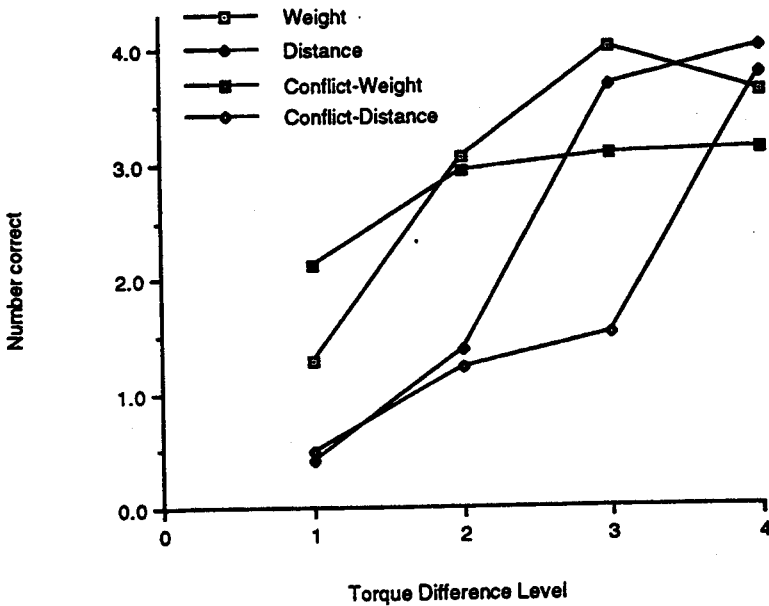


FIG. 5.6. Differences in accuracy for balance scale problem types at different levels of torque difference for networks with prestructured weight experience.

consisting of problems chosen from a unique torque difference level, that were used in the cross-sectional analysis of the torque difference effect. At each of the four levels of torque difference, there were four problems representing each of Siegler's six problem types. The *tss* error scores from each output epoch sampling demonstrated strong negative linear trends. Although a measure of this sort is not strictly applicable to human data, this trend demonstrated that the larger the problem's torque difference, the smaller the error observed.

In summary, the prestructured weight dimension model demonstrated that all four of Siegler's rules do indeed fall naturally out of a cascade-correlation network that starts off from an initial state structured to specifically process weight but not distance information. Moreover, networks in the current simulation demonstrated some of the finer subtleties of human performance such as U-shaped development curves on conflict-weight problems and the torque difference effect.

Balance Scale Conclusions

The cascade-correlation balance scale models reported in this chapter capture the phenomena in children's balance scale data, including orderly stage progression and the torque difference effect. Both cascade-correlation models achieved a consistent level of Stage 4 performance, and at least one of the models also demonstrated U-shaped error performance for conflict-weight problems. Furthermore, both models accomplished their successes without the need to prespecify the network topology, or to assume separate processing structures for the weight and distance dimensions. It would appear that a generative connectionist learning algorithm is not only capable of implementing successful models of cognitive development on the balance scale task, but does so with fewer assumptions and a greater level of success than do previous methods. Although the use of a domain-general learning algorithm was crucial, both models required domain-specific constraints in terms of either the structure of the environment or the initial placement of the network in connection space. There is not yet definitive evidence for differentiating the better set of modeling assumptions, although both models do make predictions about the learning environment. One model predicts that there is an environmental bias, whereas the other does not. Both models predict that Stage 4 performance does not necessarily require explicit knowledge of the torque rule, and that this competence can be achieved by adequate exposure to the problem domain.

The cascade-correlation learning architecture (and connectionism in general) greatly constrains the range of successful implementations of the balance scale task. Only a few specific sets of underlying assumptions yield the desired longitudinal performance. It will be of telling interest, for

purposes of model and theory building, to discover what characteristics of successful connectionist implementations correspond with human data. Our modeling efforts have captured a number of key phenomena observed in the human cross-sectional data, including some aspects of performance that other models failed to achieve. Accurate models of this nature can provide a means of investigating longitudinal properties that are difficult to reveal solely via the typical cross-sectional investigations of children.

CONCEPTS OF POTENCY AND RESISTANCE IN CAUSAL PREDICTION

Accurately predicting the magnitude of a physical effect requires the integration of information regarding potency of the cause and resistance to the effect's occurrence. In some physical systems, potency and resistance are combined in a subtractive manner ($p - r$) to produce the effect, whereas in others, they are combined by division ($p \div r$).

Psychology of Potency and Resistance

Past research on the development of these concepts revealed a number of psychological regularities. Zelazo and Shultz (1989) conducted a study with two pieces of physical apparatus, one of which combined potency and resistance using a subtraction rule (a two-tray balance scale) and the other combining potency and resistance using a division rule (a ramp). From 1 to 6 equal weights of identical appearance could be placed on the *potency* tray of the balance, and from 1 to 6 identical weights could be placed on the *resistance* tray of the balance. The magnitude of effect was indicated by the degree of deflection of a dial on the face of the scale. The six levels of potency and six levels of resistance generated 36 possible patterns. In an analogous way, from one to six wooden blocks of identical appearance could be placed at the top or bottom of the ramp. The number of blocks placed at the top constituted the manipulation of potency; the number of blocks placed at the bottom constituted the manipulation of resistance. Magnitude of effect was the distance traveled by the leading edge of the leading block at the bottom of the ramp after the collision caused by the release of the blocks at the top. The six levels of potency and six of resistance for the ramp generated 36 effect size patterns.

With increasing age, children showed an increase in the number of levels of potency and resistance used and gradual convergence on the correct rule. The subtraction rule was acquired earlier than the division rule, and there was temporary overgeneralization of subtraction to division problems.

Potency and Resistance Simulations

We were able to simulate these psychological regularities in cascade-correlation networks (Shultz, Zelazo, & Strigler, 1991). As in the psychological research, 72 problems were created by combining six levels of potency and six levels of resistance with two different combination rules, subtraction and division. These problems were the training patterns for some of the network simulations. Potency and resistance were coded in a variety of different ways in different simulations, but we focus here on what we called *Gaussian* coding. The six amounts of potency or resistance were coded over six input units such that the n th unit received an input value of 3 and the two surrounding units an input value of 1 each. Other input units in the same bank received an input of 0. Thus, the input unit activations approximated a Gaussian distribution. There were six such units for potency and six for resistance. In addition, there was an apparatus unit that was coded as 0 for subtraction and 1 for division. All inputs and the bias unit were fully connected to a single output with a linear activation function that represented the magnitude of effect, scaled to fall between 0 and 1.

Before each epoch of training, the network was tested with the 36 subtraction training patterns, the 36 division training patterns, and 36 non-trained patterns in which the subtraction outputs were associated with an apparatus unit coded for division. Comparing error scores on the first two sets of patterns provided a measure of how well the network was learning the correct subtraction and division rules, respectively. Comparing error scores on the second and third pattern sets enabled an assessment of whether or not the network might be using subtraction to solve the division problems.

Learning and overgeneralization effects in one network are portrayed in Figs. 5.7 and 5.8, respectively. Fig. 5.7 shows earlier and deeper learning of subtraction than of division. Fig. 5.8 shows a sizable overgeneralization effect in which, just prior to asymptotic performance, the error for a subtraction rule on the division problems is lower than that for a division rule on five epochs. All of the Gaussian coded networks showed these general patterns, although not at precisely the same epochs.

To examine the ability of these networks to simulate the gradual increase in levels of potency and resistance, a separate simulation was run in which network predictions were generated every fourth epoch (up to 20 epochs) for all of the 72 training patterns. These predictions were analyzed in the same manner as for human subjects to obtain the number of levels of potency and resistance employed on both subtraction and division problems (Zelazo & Shultz, 1989). The results, plotted in Fig. 9 for each of four apparatus and potency versus resistance combinations, reveal a steady increase in the numbers of levels used. For comparison, the mean numbers of levels used by human subjects are listed in the upper left corner of Fig. 5.9.

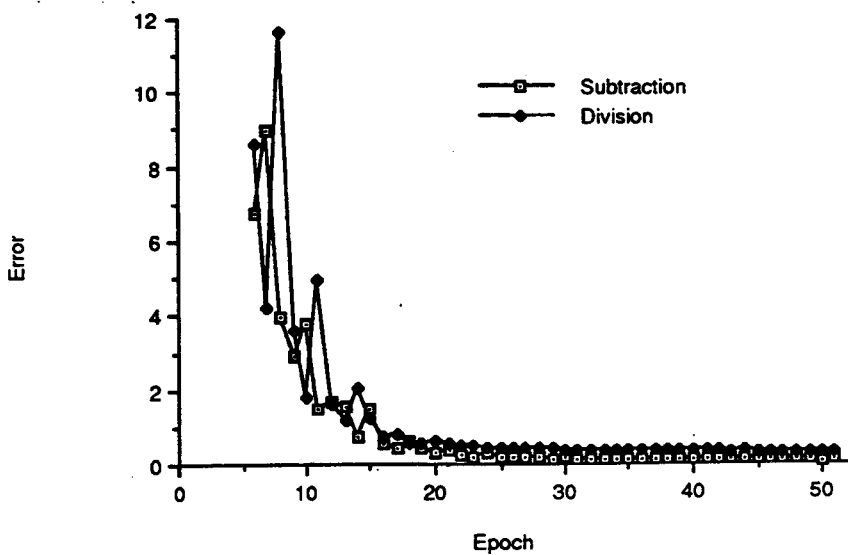


FIG. 5.7. Learning of subtraction and division in a potency and resistance network.

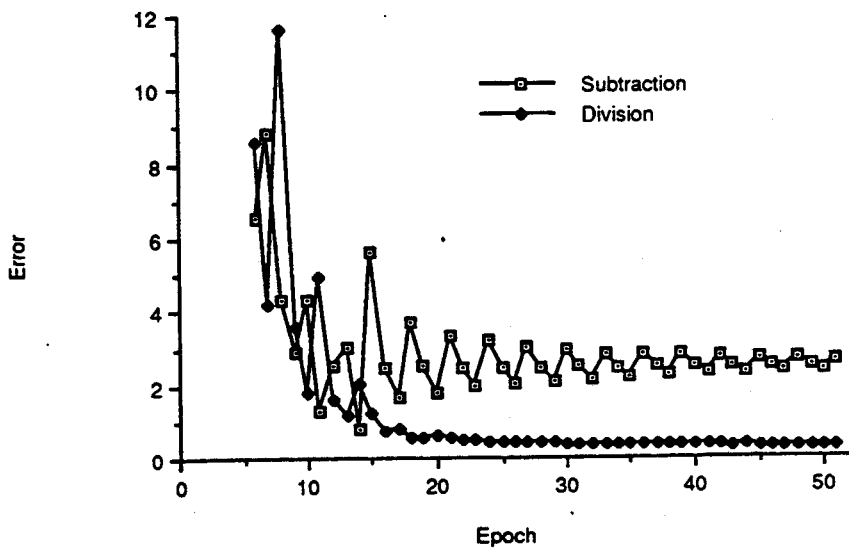


FIG. 5.8. Temporary overgeneralization of subtraction to division problems in a potency and resistance network.

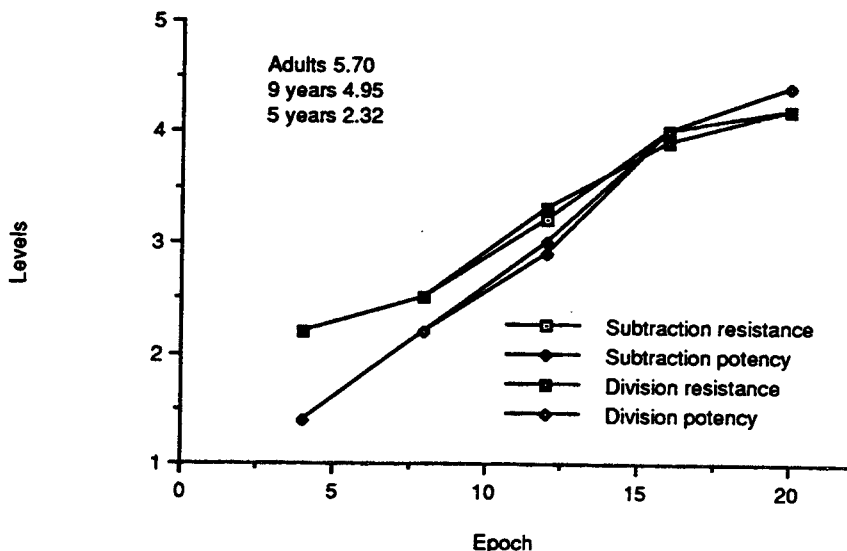


FIG. 5.9. Mean numbers of levels of potency and resistance used over epochs.

One concern with networks that use distributed binary input codings is that, because they employ so many input units and consequently have so many weights, they might be memorizing the training patterns rather than abstracting useful generalizations about them. Generalization ability was assessed by training the networks with a randomly selected two thirds of the training patterns and testing on both the training patterns and the one third nontrained patterns at each epoch. Mean error for 10 networks at each epoch for training and test patterns is presented in Fig. 5.10. The results reveal that error for the test patterns decreases with that for the training patterns, indicating good generalization.

All of these results held up with a variety of different input coding techniques, except that coding potency and resistance with integer values precluded a gradual construction of these dimensions. All of the cascade-correlation networks we tried reached an asymptote error close to the adult error of 1.08, computed in the same way as network error. None of the networks recruited any hidden units to reach this level of performance.

We had more difficulty simulating these effects with back-propagation networks, whether or not they were equipped with hidden units. This is probably due to the fact that units with sigmoid (S-shaped) activation functions are somewhat unstable in their middle range in that a small change in net input can produce a large change in activation. Using linear output units was avoided because excessive positive feedback can occasionally lead to diverging output activations. Because cascade-correlation switches to a hidden unit recruitment mode if a fixed number of epochs has

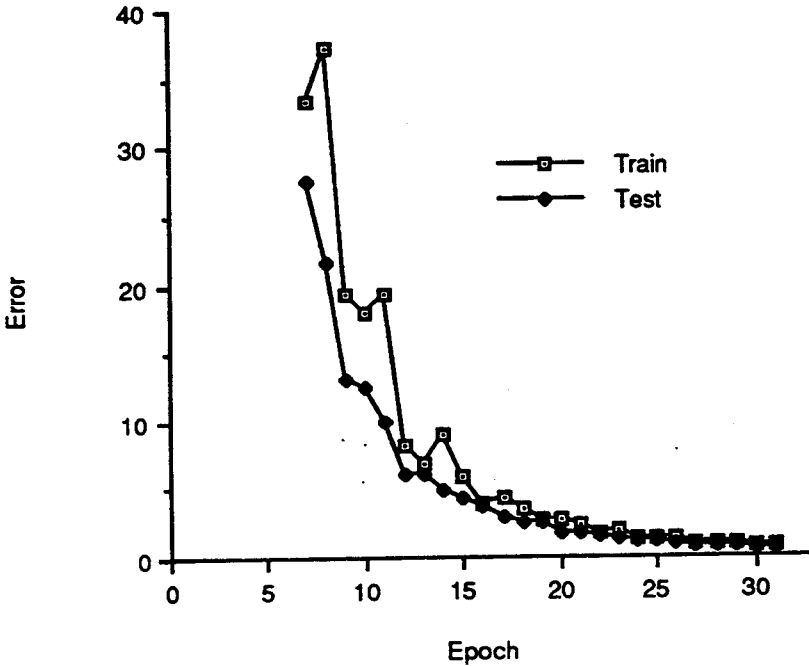


FIG. 5.10. Mean error at each epoch for training and test patterns on potency and resistance problems.

passed, it can escape from such cases. Thus, it is relatively safe to use linear output units in the cascade-correlation architecture. The option of linear output units in cascade-correlation thus provided an unexpected advantage. Further experimentation revealed that it was possible to simulate the human regularities with nonlinear output units, but these effects are more robust with linear outputs.

Potency and Resistance Conclusions

Network simulations of the concepts of potency and resistance covered all of the phenomena found in the effect size predictions of children. These included an increase in the number of levels of potency and resistance used and convergence on the correct rules as learning progressed, early acquisition of the subtraction rule, and temporary overgeneralization of subtraction to division problems. All but the first of these phenomena would appear to be due to the domain-general characteristics of cascade-correlation learning, particularly to the use of an activation function that sums the inputs to a receiving unit. An additive activation function makes

it easier to learn subtraction than division. Visual inspection of network weights indicated the networks were using the apparatus unit to dampen down predictions for division problems. Because the networks were using the same set of weights to solve both subtraction and division problems, the subtraction rule emerged first and temporarily generalized to division before the weights were sufficiently adjusted to lower the division errors.

The gradual increase in the number of levels of potency and resistance used to predict effect size is due to a domain-specific constraint of input encoding that does not inherently represent quantitative dimensions. The Gaussian coding technique used here is one of a number of coding techniques that fail to provide quantitative dimensionality. Initially, the network knows nothing about the relations among adjacent input units. As learning progresses, the network constructs the dimensions of potency and resistance. In contrast, coding inputs as integers on single units fails to generate an increase in levels used, because the quantitative dimensions are explicitly provided to the network.

SERIATION

Piaget and his colleagues (Piaget, 1965; Piaget & Inhelder, 1973) developed the seriation task in order to demonstrate the presence of developmental stages in children's transitive reasoning. Because the results initially seemed clear and replicable, a number of production rule models of children's seriation were published (Baylor, Gascon, Lemoyne, & Pothier, 1973; Young, 1976). However, these models remained structural descriptions in the sense that they depicted behaviors at particular stages and did not implement any transition mechanisms.

In this section we describe a cascade-correlation model of the development of children's seriation abilities. A more detailed account can be found in Mareschal and Shultz (1993). We begin with a brief review of the relevant psychological literature. Then, the network architecture is described. Finally, the model's performance is presented and evaluated.

Psychology of Seriation

Piaget (1965) reported that children's ability to seriate (sort) a set of objects along a specified dimension evolved through four successive stages of performance. Children in the first stage made no real effort to order the objects and either moved them around at random or presented the objects in the same order as they had found them. Children in the second stage were able to correctly order subsets of the set of objects, but did not extend this order over the whole set. Hence, they would construct successive ordered

subsets of two, three, or four objects. This led to a variety of characteristic outcomes including ordered pairs of triplets, a series that first rises then falls off, and correct seriation of the first few elements followed by inability to continue the series appropriately. Children in the third stage sorted the complete set of items, but only by using an empirical trial and error method with many self-corrections. Finally, children classified as being in the fourth stage constructed an ordered set quickly and efficiently by applying what appeared to be a systematic strategy. Piaget labeled this strategy the *operational method*. It consisted in selecting the smallest, as yet unordered, element and moving it into its correct place in the series.

Rule-based models (Baylor et al., 1973; Young, 1976) succeeded in capturing the systematicity of performance at each of these four stages. However, in depth protocol analyses revealed that seriation was far more flexible than Piaget suggested (Young, 1976). Random selection strategies were observed in children of all ages, including those well into the fourth, operational stage (Kingma, 1982).

Moreover, perceptual factors were found to influence children's performance. Piaget (1965) noticed that if the differences between elements were too large, then Stage 3 seriators would artificially be promoted to Stage 4 because the empirical method they use would be efficient given the high perceptual salience of the dimension generating the order. Conversely, if differences between objects become sufficiently small, seriation performance deteriorates (Elkind, 1964; Kingma, 1984). Also, Koslowski (1980) showed that Stage 1 seriators could be made to seriate at a Stage 4 level if given an abbreviated task with few items. She suggested that these children possessed the requisite seriation skills and that there is development in the precision with which these skills are applied.

The production rule models not only fail to capture stage transitions but do not address the perceptual saliency issues. A connectionist approach suggests an alternative modeling solution because (as illustrated throughout this chapter) it can capture rulelike behavior and perceptual effects without sacrificing flexibility.

The Seriation Model

We adopted an approach first suggested by Young (1976), who decomposed seriation into a succession of independent moves based on immediate perceptual features. Similarly, our cascade-correlation model is designed to respond to independent arrays with an appropriate move. Once a move has been computed by the network model, it is carried out by supporting software, and the network is then presented with the resulting array. Iterating this procedure may ultimately produce an ordered array.

The network was given the dual task of identifying which item should be

moved and where it should be moved, according to Piaget's operational method. We opted for a modular solution because simulations using a single homogeneous network failed to demonstrate psychologically realistic performance. A number of researchers argued for modular task decomposition (Jacobs, Jordan, & Barto, 1991; Minsky, 1986). Modular architectures were claimed to increase both learning speed and generalizability (Jacobs et al., 1991).

The two modules consist of two simultaneously but independently trained networks, as shown in Fig. 5.11. The *which* network is trained to identify which item should be moved when presented with an array. The *where* network is trained to identify where an item should be moved to when presented with the same array. In both cases, targets are defined as the move dictated by Piaget's operational method. As noted earlier, this involved selecting the smallest unordered stick and placing it in its correct position. The modules are independent because the weight updates within each network are based solely on the error arising within that network. Each network has no information concerning the performance of its counterpart. Because each move requires integrating responses across modules, the macroscopic behavior of the model as a whole results from the interaction of the developmental states of each independent module. Thus, systematic errors may arise when one module lags behind the other in performance.

The model was trained to seriate an array of six items as follows. Each item had a unique value determined by an integer ranging from 1 to 6. The location of the item was spatially coded on a bank of six linear input units. Thus, a completely ordered array was coded by a 1 in the first unit, a 2 in the second unit, and so forth, with a 6 in the sixth unit. The output (whether *which* or *where*) was coded on a bank of six sigmoid units in which the unit coding the correct position is turned on and all others are turned off. The actual response was determined by selecting the output unit with the highest activation.

Pilot studies revealed that the model was sensitive to the disorder of the arrays presented. Disorder was quantified as the sum squared distance (d^2) from the target ordered array. Input patterns were classed as being distant from the solution if d^2 was greater than 20 and near to the solution if d^2 was less than or equal to 20. Of the total 720 possible six-element patterns, 79% are distant patterns and 21% are near patterns. A biased training set was constructed by randomly sampling 50 distant and 50 near patterns. In order to capture the fact that even very young children can successfully order sets of 3 elements, we included 20 3-element series in the training set.

The model's performance was evaluated in two ways. Generalization was tested by evaluating whether or not the model produced the correct move for all of the possible six-element series. Its seriation stage performance was

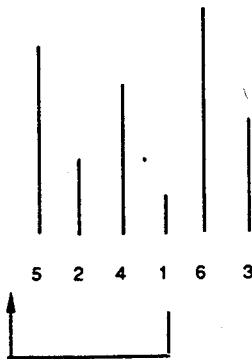
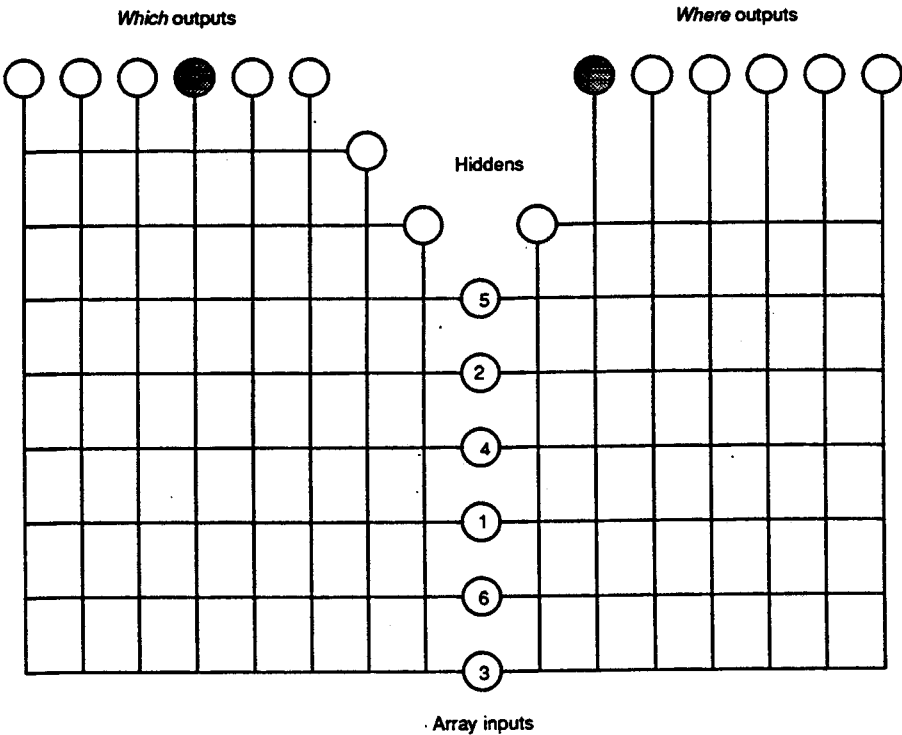


FIG. 5.11. Modularized seriation networks. The row of vertical units contains the input units on which is represented the current status of the array to be seriated. The input feeds independently to the left-hand and the right-hand modules. The left-hand side is devoted to computing which item in the array should be moved, and the right-hand side is devoted to computing where the item should be moved to. In this case, the *which* module has two hidden units and the *where* module has one hidden unit. Inputs and target outputs are shown for the array {5 2 4 1 6 3}. The stick in the fourth position should be moved to the first position of the array.

evaluated by presenting a test pattern to the network.³ The state resulting from each move was cycled back as the next input. The cycling process continued until the presence of a loop was detected. From the resulting trace of arrays, the model was then classified as being in one of the four developmental stages.

Stage classification can be somewhat of a problem, not just for networks, but also for children (Kingma, 1982). In particular, it is not always clear how to differentiate between Stage 3 and Stage 4 seriators. In our simulations, Stages 1 and 2 are diagnosed as described by Piaget. To distinguish between Stage 3 (empirical) and Stage 4 (operational) seriators, both the procedure used and the number of self-corrections criteria are simultaneously applied. A network is classified as Stage 4 if it correctly constructs a series according to the operational method with at most one error from which it continues using the same operational method, or if it seriates in the same or fewer moves than required by the operational method. It is classified as Stage 3 if it constructs a completed series in any other way. Under these conditions, networks typically exhibited a succession of all four stages in the correct order.

To test whether or not these networks responded to perceptual variations in the same way as children do, we ran three additional conditions. In these, the training set consisted only of the 100 6-element arrays. The three conditions differed only in the size of differences between successive elements of the ordered set: 1.0, 0.5, 0.25. The proportions out of 20 models able to complete a full sort by the end of training were 0.85, 0.55, 0.15 respectively. Thus, as with the children, the more easily the stick sizes can be distinguished, the better the seriation performance. Furthermore, 85% of the models trained with 1.0 size differences were diagnosed at Stage 4, as compared to only 25% of the nets trained with 0.5 size differences. This supports Piaget's (1965) claim that Stage 3 performers would be classed at Stage 4 as size differences between sticks increase.

Inspection of Hinton diagrams generated at epochs representing consistent stage behavior revealed no drastic differences in weights between adjacent stages. Instead, stage differences were marked by rather small modifications in the size of weights. The Hinton diagrams also revealed that the development of seriation ability began by adjusting weights leading to those units dealing with the short end of the series and was progressively extended along the length of the series until appropriate weights were found for the larger end of the series.

Finally, the disorder of the array positively predicted the model's ability to identify the correct move in the generalization test. In a study carried out on children aged 4 to 7 years, we found that disorder was similarly related

³Following Retschitzki (1978), we used the {5 2 4 1 6 3} array.

to children's decisions as to whether a series was completed, or still needed some sorting (Mareschal, 1992).

Seriation Conclusions

Cascade-correlation nets captured progression through the four seriation stages, as well as a perceptual effect based on the differential sizes of the items to be sorted. Another perceptual effect on array disorder was noted, and some supporting psychological evidence was provided by studies of children who showed more difficulty with less disordered arrays.

A number of domain-specific constraints were required to capture seriation stages, including a modularization of the task into *which* and *where* subnets and a slight environmental bias in favor of smaller and less disordered arrays. Both of these biases seem reasonable. Smaller arrays are probably more common to the young child's experience than large ones, and less disordered arrays are more likely to function as a cue to sorting attempts than are highly disordered arrays. In any case, these biases could be considered as predictions for the young child's environment. Task modularity could also conceivably be assessed through psychological research that concentrated on possible dissociations between selection and insertion abilities.

Perceptual effects in seriation can be attributed to the domain-general characteristics of the learning algorithm. As with the balance scale task, more distinctive quantitative inputs naturally result in clearer activation signals downstream and more decisive moves; inputs to hidden and output units are a function of the activation values of sending units.⁴

DISTANCE, TIME, AND VELOCITY

Recently we began to extend our research from tasks that involve the integration of two physical dimensions (e.g., weight and distance in the

⁴In chapter 8 (this volume), Klahr criticizes our seriation model for not exhibiting the multiple strategies typical of children. However, our model is motivated by the fact that it is the rule-based models that are too rigid to account for both individual variation among children and the variation across seriation problems due to perceptual effects on the seriation task (as their authors freely admit). In contrast, even though we constructed our training patterns with the single rule specified by Piaget (move the smallest unordered stick to its correct position), we found considerable variation in performance among nets and problems (Mareschal, 1992). A variety of behaviors fell under various stage diagnoses, and there was sufficient variation to account for the various perceptual effects. This is the only existing seriation model to spontaneously generate variation, and much of that variation does correspond to variation in children.

balance scale task, and potency and resistance in causal prediction) to those involving three dimensions. The example we cover here is the integration of distance, time, and velocity concepts. In classical physics, distance is defined as $d = t \times v$, time as $t = d \div v$, and velocity as $v = d \div t$, where d is distance, t is time, and v is velocity.

Psychology of Distance, Time, and Velocity Concepts

Piaget (1969, 1970) investigated the development of these concepts after Einstein had inquired about the nature of children's understanding of time and velocity. Piaget's research led him to conclude that the acquisition of these concepts occurred in three stages. At 4 or 5 years of age, intuitive notions emerge. For example, children's early concept of distance traveled is in terms of the stopping point of an object rather than the interval between starting and stopping points. These early intuitions are followed by an intermediary stage and finally, the adultlike concepts emerge at approximately 8 or 9 years of age. In response to Einstein's inquiry, Piaget concluded that although an intuitive notion of velocity existed independent of time, the notion of time was dependent on the child's notion of velocity at most ages. Thus, children's early understanding was more akin to relativistic concepts of time and velocity.

Siegler and Richards (1979) addressed several methodological difficulties in Piaget's work, including the use of tasks that were not necessarily comparable across concepts. Siegler and Richards presented children with two toy trains running along parallel tracks and asked them to judge which train either traveled for the longer time, the greater distance, or faster. They hypothesized three rules based on Piaget's work. Children using Rule 1 would make their judgments based on the stopping points of the trains. Those using Rule 2 would also consider starting points when the trains stopped at the same point. Finally, children using Rule 3 would solve the problems correctly.

Siegler and Richards used a rule assessment methodology similar to that employed with the balance scale. Their results indicated that 5-year-old children used Rule 1 on all three tasks, whereas adults used Rule 3. In between these two age groups, children often confused velocity and distance, distance and time, and time and velocity. In addition, children understood distance and velocity concepts before time concepts.

Levin (1977) examined children's understanding of time and argued that in tasks used by Piaget and Siegler and Richards, distance and velocity information served as interfering cues with children's understanding of time. Moreover, Levin (1979) argued that cues logically unrelated to time interfere in a similar manner.

Wilkening (1981) made a similar argument, suggesting that research by Piaget and others appeared to have tested the child's ability to ignore rather

than integrate dimensional cues. For example, to judge which train traveled the greater distance, the child simply had to compare the distance of the two trains and ignore their times and velocities.

Within the framework of Anderson's (1974, 1991) Information Integration Theory, and its assessment methodology, *functional measurement*, Wilkening (1981) designed new tasks in which values on two dimensions were given and the value of the third dimension had to be inferred by the child. For example, in a distance-inference task, children were shown an apparatus that had, at one end of a footbridge, a dog and several other animals that were said to be frightened of the dog. The children were told that the other animals would run along the bridge as soon as the dog began to bark and would stop when the barking ceased. The task involved determining how far each animal would run. Thus, the children were given the characteristic velocity of the animals and the time they ran (the duration of barking), and asked to infer the distance they would run.

Wilkening studied the performance of three age groups: 5-year-olds, 10-year-olds, and adults. The findings included the following: (a) in the distance-inference task, all age groups used the correct multiplication rule; (b) in a time-inference task, 10-year-olds and adults employed the correct division rule, whereas 5-year-olds used a subtraction rule, $t = d - v$; (c) in a velocity-inference task, the two older age groups used a subtraction rule, $v = d - t$, and the 5-year-olds used an identity rule, $v = d$.

Wilkening concluded that young children did have the ability to integrate these dimensions. However, he was unwilling to make comparative claims about the developmental rates of the three concepts because it appeared that the subjects had differing memory demands across the three tasks. For example, in the distance task, subjects of all age groups used an eye-movement strategy in which they appeared to "follow" the imaginary animal as it ran across the footbridge.

In a follow-up study, Wilkening (1982) attempted to increase the memory demands of the distance task by presenting the time information (barking) before the velocity information (animal identity) and lessen the memory demands of the velocity task by visually presenting the time information. The modifications partially supported his hypothesis in that 5-year-olds were observed to use an additive rule ($d = t + v$) in the distance task. However, the results of the velocity task remained unchanged. Thus, it remains to be seen whether or not the mastery of time before velocity concepts is an accurate description of the developmental course or a memory artifact of Wilkening's tasks.

Simulating the Acquisition of Distance, Time, and Velocity Concepts

We followed Wilkening's example by creating input patterns that included information about two dimensions and having the network predict, as

output, the value of the third dimension. We chose dimensional values ranging from 1 to 5 as our input values. The dimension that was to be predicted received an input of 0. We crossed five levels of distance, time, and velocity respectively to obtain 75 input patterns. There were 25 instances of each of the three inference pattern types: distance, time, and velocity.

The initial network topology consisted of three banks of input units, one bank each for distance, time, and velocity information, connected to one linear output unit. We used what we call *n*th encoding for the dimensional values on the input units. In *n*th encoding, a value *n* is represented by assigning an input value of 1 to the *n*th input unit and 0 to all other units. Thus, to encode the 3 dimensional values having a range of 1 to 5, a total of 15 input units were used—5 units for each dimension. For a given inference pattern, one input bank would receive activations of 0 on all five of its inputs, indicating it was unknown. With respect to the other two input banks, the appropriate unit would receive an activation of 1 corresponding to its dimensional value and 0 on the other units within the bank.

Training And Testing. At each epoch of training, all 75 inference problems were presented to the network. Thirty networks were trained for a maximum of 1,500 epochs. Every fifth epoch, the net was tested to obtain the relevant information necessary to assess its performance.

In analyzing a network's performance, we are not so much interested in whether or not the network can accurately predict, for example, a given velocity from time and distance information. Rather, we are interested in what sort of rule best captures the network's predictions over each of the three problem types. Thus, we look at correlations between the network's responses and those predicted by various plausible rules such as identity ($v = d$, or $v = t$), addition ($v = d + t$, or $v = d - t$), or multiplication ($v = d \times t$, $v = t \div d$, or $v = d \div t$) rules. In this way, we investigated networks' ability to capture the stage progressions observed by Wilkening with children. In order to capture consistent network performance, a given rule had to correlate positively with network responses, account for more than 50% of the variance in network responses, and account for more of that variance than other rules did.

Results. A stage-by-epoch plot of a typical network based on stage onset and offset and hidden unit recruitment is shown in Fig. 5.12. All 30 networks demonstrated a similar developmental sequence. As can be seen, early in training, before the recruitment of any hidden units, time and velocity identity stages ($t = d$ and $v = d$) were observed. On average, these identity stages began together and lasted for the same length of time. The time and velocity identity rules were strong predictors of the networks' responses, accounting for over 90% of the variance in predictions. During

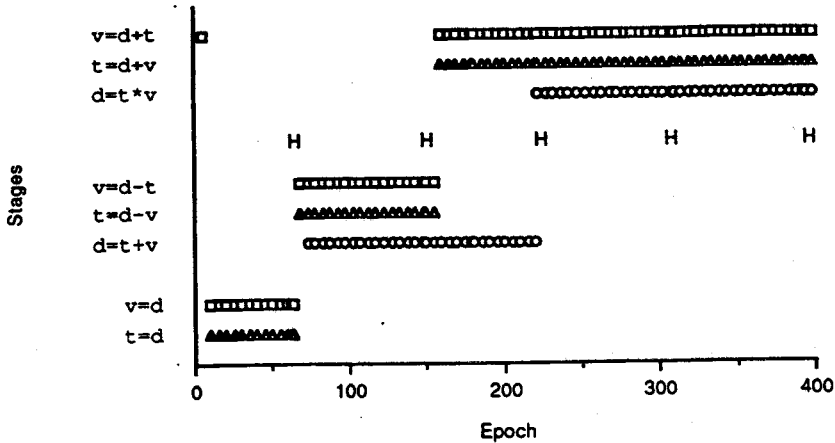


FIG. 5.12. Rule diagnosis results for one network showing the progression of stages over epochs for distance, time, and velocity inferences. Each *H* marks the recruitment of a hidden unit.

this same period, the networks' responses to distance-inference patterns was not captured by any of the rules that were tested.

After the first hidden unit had been recruited, additive stages were observed with respect to all three inference types. Distance, time, and velocity inferences were captured by the additive rules $d = t + v$, $t = d - v$, and $v = d - t$, respectively. Although all three of these stages began at approximately the same epoch, the distance additive stage was typically longer. These 3 additive rules were good predictors of the networks' responses, accounting for over 80% of the variance in predictions.

Multiplicative stages of time ($t = d \div v$) and velocity ($v = d \div t$) inferences began just after the second hidden unit was recruited. On average, the multiplicative stage of distance inferences ($d = t \times v$) began after the third hidden unit was recruited. All three defining multiplicative rules of the stages eventually reached a maximum r^2 of 1.00. This occurred earlier for time and velocity inference patterns than for distance inferences. See Buckingham and Shultz (1994) for a more detailed presentation of these results.

Analysis of Hinton diagrams revealed that the first hidden unit distinguished distance inference patterns from time and velocity inference patterns. Typically, weights from the time and velocity input banks were of the same sign and opposite in sign to weights from the distance input bank. As a result, when a distance inference pattern was presented, the time and velocity inputs augmented each other. In contrast, when a time or velocity inference pattern was presented, the distance input was counteracted by the velocity or time input. Unfortunately, Hinton analysis of the second and third hidden units were less revealing. However, given the relatively abrupt

transition after the installation of these hidden units to multiplicative stages, the need for increased nonlinearity seems evident.

Conclusions on Distance, Time, and Velocity

Our simulation results differ only marginally from Wilkening's (1981) human results. The distance developmental sequence observed in network performance is the same as the one found by Wilkening—a progression from an additive rule to correct integration based on multiplication. With respect to the time and velocity developmental sequences, the networks began with an identity rule, progressed to an additive rule, and then finished with a correct multiplicative rule. Wilkening's human subjects did the same, except that they showed no identity rule for time inferences and failed to reach the correct multiplicative rule for velocity inferences. Our network results predict these "missing" stages for younger and older (more experienced) subjects than those used by Wilkening.

Identity, additive, and multiplicative stages in network performance emerged from the domain-general constraints of the cascade-correlation learning algorithm. That is, identity rules arise from the limited computational abilities of cascade-correlation's initial perceptron topology. Although simple weight adjustment is sufficient to decrease a substantial proportion of the sum of squared error of the various inference patterns, it is insufficient to allow the emergence of performance characterized by more advanced rules. After the recruitment of a single hidden unit, more complex performance emerges that can be characterized by additive rules in which two known values are added or subtracted to predict a third value. Finally, performance characterized by the correct multiplicative rules requires further computational nonlinearity provided by the recruitment of additional hidden units.

Developmental performance, characterized by specific algebraic rules, can be simulated in a network that learns by simple weight adjustment and hidden unit recruitment. The progression from linear to nonlinear rules parallels the potency and resistance simulations presented earlier. Such progressions occur naturally in cascade-correlation nets because they employ units with an additive activation function.

ACQUISITION OF PERSONAL PRONOUNS

Consider the following interchange:

Father (to daughter): "OK, Jane, let's practice our pronouns."
Jane (enthusiastically): "Okay, dad."

Father (pointing to himself): "Me" (and then to daughter): "You!"
Jane (pointing to herself gleefully): "You!"
Father (pointing to himself): "No, no. Me!"
Jane (confused, points to father): "Me!"
Father (frustrated after several failed attempts): "OK, me Daddy, you Jane."

This fictional dialogue between a father and his young daughter is meant to illustrate two problems a child faces when trying to learn the correct use of personal pronouns such as *me* and *you*. First, the referent of *me* and *you* is not fixed but shifts with conversational role. For example, when a child's father and mother talk to each other, both refer to themselves as *me* and to the other as *you*. Thus, the referent of *me* and *you* depends on who is speaking and who is being addressed. Second, the model for correct use of personal pronouns is not ordinarily given in speech addressed directly to the child. As just demonstrated, when a father addresses his child, he refers to himself as *me* and to the child as *you*. If Jane were to imitate what she heard, she would incorrectly refer to herself as *you* and to her father as *me*. Such errors have been called *reversal errors* because the child reverses the correct use of the pronouns.

Psychology of Personal Pronouns

Given that the task of learning personal pronouns is so complex, it is remarkable that most children master their correct use by 3 years of age (Clark, 1978). Perhaps even more remarkable is the fact that the majority of children do so without reversal errors (Charney, 1980b; Chiat, 1981). However, some children, like the fictional Jane, do make reversal errors, and such errors can often persist for months (Clark, 1978; Oshima-Takane, 1992; Schiff-Meyers, 1983).

Theories of personal pronoun acquisition can be placed in one of two categories: those focusing on children's correct performance, and those focusing on the errors children make. Within the former category, research focused on speech roles (Shipley & Shipley, 1969) and imitation without understanding (Charney, 1980b). Examples of theories focusing on errors include children's inability to distinguish self from other (Bettleheim, 1967; Charney, 1980a) and the interpretation of pronouns as names, in which the first person pronoun equals the name of the parent and the second person pronoun equals the name of the child (Clark, 1978). Regardless of the focus, these theories can account for only part of the picture. Focusing on errors does not explain how the majority of children master personal pronouns without error, and focusing on correctness fails to explain why some children make persistent reversal errors.

In an attempt to understand the variation in children's pronoun errors, Oshima-Takane (1988) hypothesized that the nature of the speech that a child hears plays a critical role. Although some researchers have argued that nonaddressed speech (in which the child is not being talked to) is unimportant in language acquisition (de Paulo & Bonvillian, 1978; Ervin-Tripp, 1971), Oshima-Takane maintained that it is this type of speech that enables the child to acquire the correct semantic rules for pronouns. The correct semantic rules specify that the first person pronoun refers to the person using it and the second person pronoun refers to the person who is addressed.

The importance of nonaddressed speech in acquiring personal pronouns was shown in both a training experiment (Oshima-Takane, 1988) and an observational study (Oshima-Takane & Derevensky, 1990). In the training experiment, 18 English-speaking, 19-month-old children and their parents participated in a pronoun game. In the nonaddressee condition, the game had two parts. In the first part, the mother pointed to herself and said *me*. Then the father said *me* pointing to himself, after which the mother pointed to the father and said "Yes, you." Immediately following this exchange, the second part of the game began. Here the mother pointed to herself and said *me* once again and then waited for the child to say *me* pointing to himself or herself. If the child said *me* the mother responded "Yes, you." If the child did not respond, the mother simply pointed to the child and said *you*. In any event, the game was then replayed.

In the addressee condition, the game involved only the second part of the nonaddressee game, except that both mother and father took turns addressing the child. The results of this experiment indicated that only the children who heard nonaddressee speech were able to use the pronouns without error. In contrast, reversal errors were common among children in the addressee condition.

In the observational study, 16 first- and secondborn children who had a sibling 1 to 4 years older were observed during free-play sessions (Oshima-Takane & Derevensky, 1990). Although the two groups did not differ on general language measures such as mean length of utterance, the secondborn children acquired personal pronouns earlier than the firstborn children. This result was predicted from the fact that secondborn children have more opportunities to hear speech not addressed to them because they hear conversations between the parent and older sibling.

We attempted to model the learning of personal pronouns in English. We used extreme versions of Oshima-Takane's (1988) training experiment, as illustrated in Fig. 5.13. In this figure, the arrows originate from the speaker, start out in the direction of the addressee, and end up pointing to the referent. We trained the network in two phases. During Phase 1, the network was trained on only parent speaking patterns, illustrated in the top

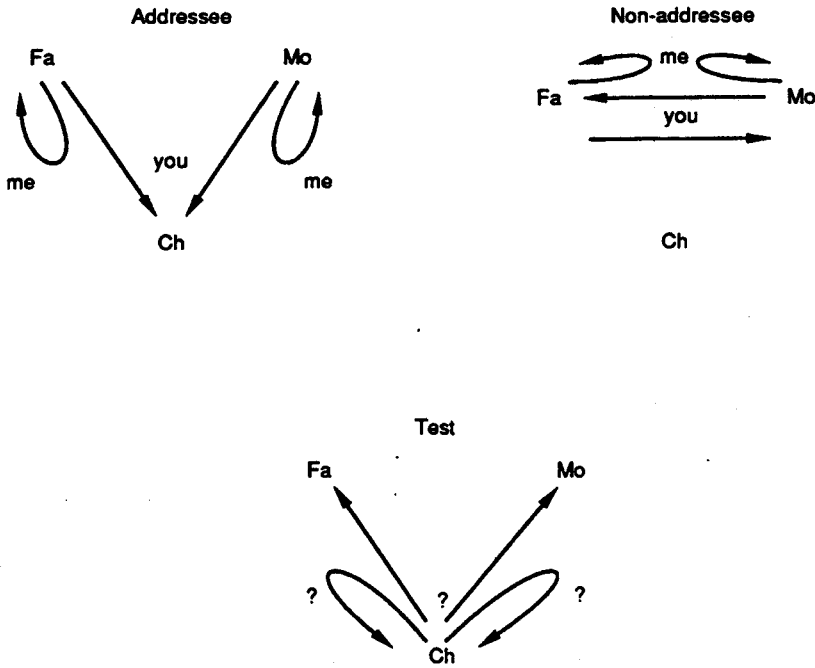


FIG. 5.13. Training patterns for the pronoun simulations. Adapted with permission from Shultz, Buckingham, and Oshima-Takane (1994).

half of Fig. 5.13, simulating the early period in which children listen to conversations without participating. In Phase 2, the network was trained on child speaking patterns, illustrated in the bottom half of Fig. 5.13, simulating the later period when children join in conversation. The critical question was how long it would take the network to learn the child speaking patterns in Phase 2 given the type of Phase 1 training it had received.

Pure Condition Simulations

In our first set of simulations (Shultz, Buckingham, & Oshima-Takane, 1994), we examined two idealized situations: one in which a child only hears speech addressed to him or her (addressee condition), and the opposite situation in which he or she only hears speech between his or her parents (nonaddressee condition).

We used four patterns in each condition. In the addressee condition, shown in the top left of Fig. 5.13, the patterns corresponded to both the mother and the father addressing the child and saying *me* and *you* appropriately. In the nonaddressee condition, shown in the top right of Fig. 5.13, the patterns corresponded to one parent addressing the other

parent and referring to the self or the other. In Phase 2, the network was trained on child speaking patterns, shown at the bottom of Fig. 5.13. There were four such patterns corresponding to the child speaking to each parent and saying *me* or *you*. Again, the critical question was how long it would take the network to learn the child speaking patterns in Phase 2 given the type of Phase 1 training it had received.

The initial network topology consisted of six input units and a bias unit fully connected to two output units (Fig. 5.14). The six input units were comprised of three pairs of units corresponding to speaker, addressee, and referent, respectively. The identities of the participants in the conversation were distributed across a pair of units as follows: 1 0 for father, 0 1 for mother, and 1 1 for child. The target values on the output units were $+0.5 - 0.5$ for *me* responses and $-0.5 +0.5$ for *you* responses.

We discovered that addressee training during Phase 1 was easy for our networks. On average, the networks required only 14 epochs to reach victory. This contrasted sharply with the length of time needed to master nonaddressee patterns, where the mean number of epochs to victory was 47 epochs. Moreover, only within the nonaddressee condition was it necessary to recruit a hidden unit. The extra computational power provided by the hidden unit was necessary to encode the shifting pronominal reference found in the nonaddressee patterns. That is, in the addressee patterns, *you* always referred to the child and *me* always referred to the mother or the father. Conversely, in nonaddressee training, the referent of *me* and *you* could be either *mother* or *father*.

Training times required for Phase 2 patterns revealed the opposite tendency. That is, networks that had received nonaddressee Phase 1 training were able to learn Phase 2 patterns very quickly ($M = 14$ epochs),

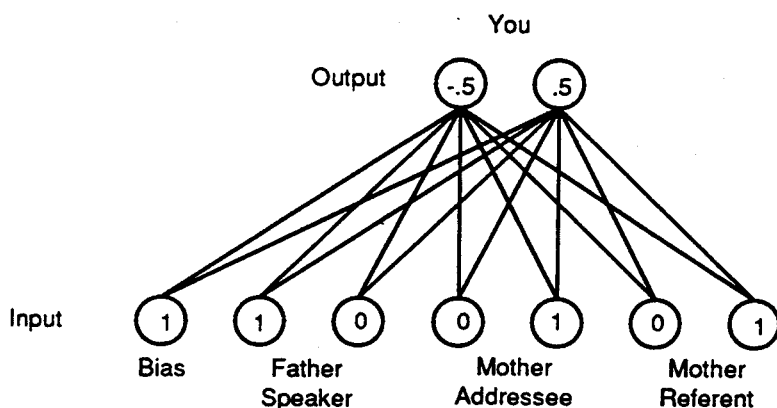


FIG. 5.14. Initial network topography and an example of input and output coding for the situation in which the father speaks, points to the mother, and says *you*.

whereas those that had received addressee training in Phase 1 required 111 Phase 2 epochs, on average. In addition, the networks in the latter condition needed to recruit a hidden unit, whereas the nonaddressee networks did not. Again, we took this as evidence that nonlinear computational power was needed to encode pronominal shifts. Because the networks that received nonaddressee training had already compensated for these shifts, they could more easily learn the child speaking patterns.

Networks in the addressee condition showed persistent reversal errors before performing correctly. Fig. 5.15 shows a plot of the output activations for a typical run in the addressee condition. The dashed lines indicate the score thresholds for positive and negative targets. For the network to be considered as using a particular pronoun, the two outputs have to be on opposite sides of these dashed lines. This network initially says *you* when it should be saying *me*: The network begins by making a reversal error. In contrast, networks in the nonaddressee condition often showed rapid correct generalization, as can be seen in Fig. 5.16. Thus, rapidly correct generalization was associated with nonaddressee training, and persistent reversal errors were associated with addressee training.

Mixed Condition Simulations

Our second pronoun simulation (Shultz, Buckingham, & Oshima-Takane, 1994) examined a more realistic learning environment. By using five

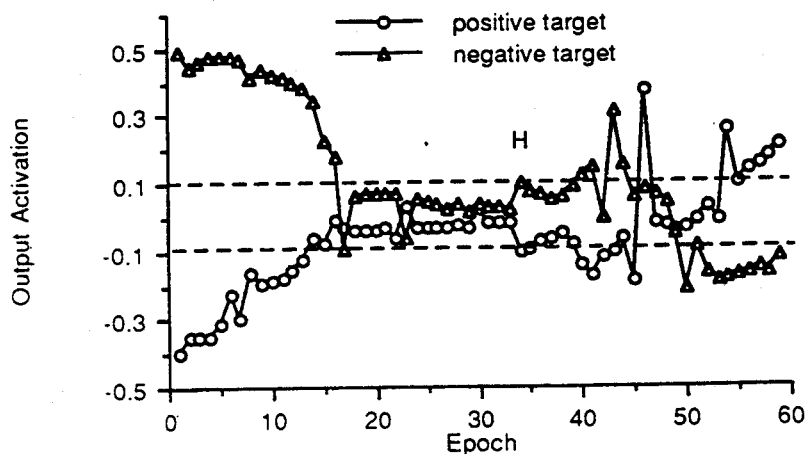


FIG. 5.15. Results for one pronoun network in the addressee condition. Output activations are plotted across epochs for the pattern in which the child is speaking to the mother and referring to self. After making persistent reversal errors, the mistake is eventually overcome following the recruitment of a hidden unit, marked by an *H*. Adapted with permission from Shultz, Buckingham, and Oshima-Takane (1994).

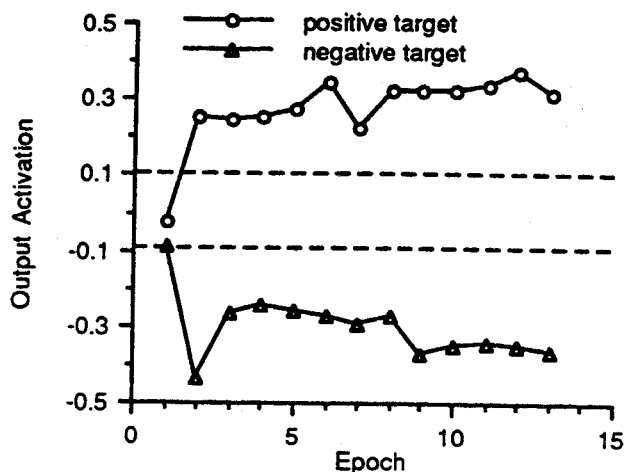


FIG. 5.16. Results for one pronoun network in the nonaddressee condition. Output activations are plotted across epochs for the pattern in which the child is addressing the father referring to self. Adapted with permission from Shultz, Buckingham, and Oshima-Takane (1994).

conditions with frequency multiples of addressee:nonaddressee patterns of 9:1, 7:3, 5:5, 3:7, and 1:9, we examined the effects of various hybrid learning environments. The 9:1 and 5:5 conditions might correspond to the linguistic environments of first- and secondborns, respectively. A firstborn is likely to hear addressee speech during the day, while one parent is away at work, and a bit of nonaddressee speech in the evening, when the working parent returns. A secondborn, in contrast, is likely to receive about equal measures of addressee and nonaddressee speech all day. The extra nonaddressee speech is provided by conversations between the caretaking parent and the older sibling.

The same network topography and training patterns used in the first experiment were used in this simulation. The only difference was in the number of exposures of a given pattern during an epoch. After the network had learned the 40 patterns of Phase 1, it was trained on the four child speaking patterns as in the previous simulation.

At first we were surprised to find that networks having more addressee than nonaddressee patterns took longer to learn the Phase 1 patterns, as is illustrated in Fig. 5.17. This was the opposite of what happened in our previous simulation. However, upon reflection, this might also be explained in terms of shifting pronominal reference. The need to encode such a shift could be temporarily masked by the frequency of addressee patterns.

The time required to learn the child speaking patterns in Phase 2 reflects what was found in the earlier simulation. As can be seen in Fig. 5.18, there is a negative linear trend associated with increased frequency of non-

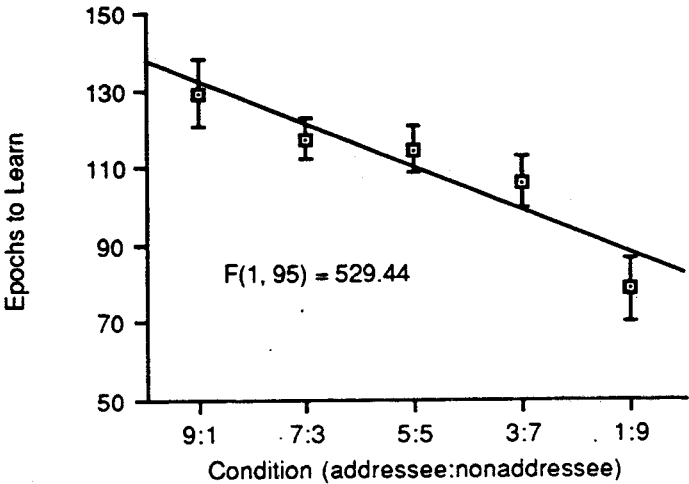


FIG. 5.17. Mean number of epochs needed to reach victory on pronouns in Phase 1 with standard deviation error bars. Adapted with permission from Shultz, Buckingham, and Oshima-Takane (1994).

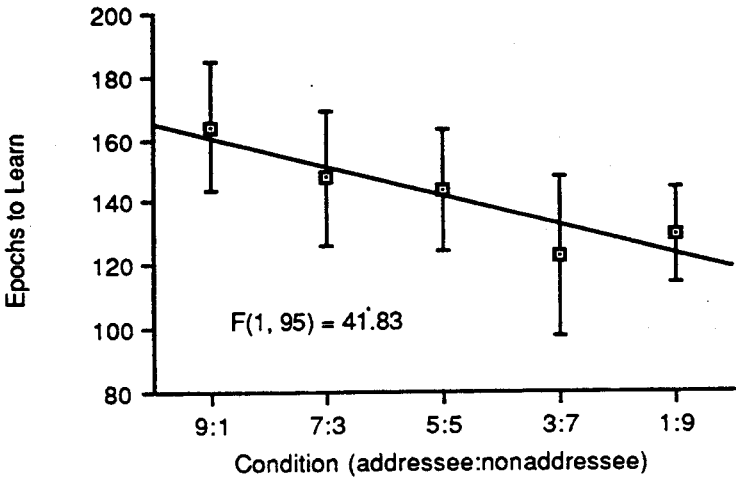


FIG. 5.18. Mean number of epochs needed to reach victory on pronouns in Phase 2 with standard deviation error bars. Adapted with permission from Shultz, Buckingham, and Oshima-Takane (1994).

addressee patterns. In other words, networks receiving more nonaddressee training learned more rapidly than networks receiving more addressee training. This supports the notion that secondborn children learn personal pronouns earlier than firstborn children because of their opportunity to hear speech not addressed to them (Oshima-Takane & Derevensky, 1990).

Adding a Self/Other Input Unit

It occurred to us that the task presented to our networks might be difficult because we assumed the network was the child, but were in no way telling it so. That is, the input patterns only indicated the identity of the individuals with respect to linguistic role (i.e., speaker, addressee, or referent). We did not tell the network that in the addressee situation the network was being talked to, and that in the nonaddressee situation it was merely listening to two other speakers. In order to be more realistic, we decided to make this information explicit to the network.

To code the identity of the network, we added a third input unit to each of the three input unit pairs. This additional unit indicated whether the individual being encoded was the child (self) or one of the parents (other). An activation of 1 was used to indicate *other*, whereas 0 indicated *self*. Otherwise, the architecture and training patterns were the same as in the first pronoun simulation.

Compared to networks in the first pronoun simulation, the self/other unit increased the rate of learning by six epochs on average in the addressee condition but had no influence on the nonaddressee condition. This latter finding was not surprising because the self/other input unit should not contribute to understanding, as it had a constant value of 1 across patterns (i.e., the parents occupied all speech roles). In the addressee condition, the added salience that parents and child were different made learning easier than the already simple addressee task.

Of more interest was whether or not the explicit information about self or other would benefit the network in Phase 2, where the child speaking patterns had to be learned. The number of epochs needed to learn in the addressee condition was cut to about one half (60 epochs) the original time. In the nonaddressee condition, there was a increase of approximately 4 epochs.

Thus, the time required for the net to learn the proper use of personal pronouns was substantially reduced in the addressee condition by the addition of self/other information. This did not alter the basic finding that nonaddressee training aided the acquisition of child speaking patterns as compared to addressee training. Moreover, error patterns found in the first simulation were unchanged in the current simulation. That is, reversal errors were still associated with addressee training, and fast and accurate generalization still reflected nonaddressee training. Thus, it appears that the findings in the previous simulations were not due to our having made the task confusing by not explicitly telling the network that it was playing the role of the child.

Pronoun Conclusions

By manipulating domain-specific constraints regarding how much addressee versus nonaddressee speech to which the network was exposed, we simulated children's performance that is marked by persistent reversal errors and correct performance. The more nonaddressee speech to which the network is exposed, the less time is required for the network to learn the correct semantic rules. An interaction between domain-specific and general constraints is responsible for correct rule use. We found that hidden units are necessary to encode pronominal shifts. In order for the networks to detect these shifts, some nonaddressee speech has to be processed and hidden units must be in place.

Use of computer simulations enabled us to look at idealized learning environments that would have been impossible to find in a child's world or to replicate within a laboratory setting. For example, we were able to have a purely nonaddressee environment where the network's attention was guaranteed without the need for participation in conversations.

The use of connectionist architectures allowed us to examine some realistic effects with respect to the frequency with which children were exposed to nonaddressee and addressee input. For example, we were able to examine the differing expected environments of first- and secondborn children. Within a symbolic framework such as Soar (Newell, 1990) such frequency effects would be difficult to obtain. For Soar, presentation of a single instance of a given type of linguistic input would typically be sufficient to create a rule to account for it. Yet, it is obvious that children hear many instances of personal pronouns as they discover the correct semantic rules for pronoun use.

We showed that prior knowledge attained from addressee and nonaddressee speech greatly affects the ability to generalize to a new situation in which the network begins to use pronouns. Our simulations, along with Oshima-Takane's (1988) findings, suggest that children acquire the correct use of personal pronouns by attending to speech that is not addressed to them. Also, it is likely that persistent reversal errors are due to children attending to speech addressed to them without having much opportunity to hear nonaddressee speech.

GENERAL CONCLUSIONS AND DISCUSSION

We reported on successful cascade-correlation simulations in several different domains of cognitive development. Most of these domains deal with the child's understanding of aspects of the physical world. In the case of a

noted benchmark task for developmental simulations, the balance scale, our network models captured the progression through four rule-based stages and the torque difference effect. The latter is a perceptual salience effect wherein problems with large absolute torque differences between one side of the balance scale and the other are easier to solve than problems with small torque differences. Our network models also captured all major psychological regularities regarding the prediction of effect magnitude from information on causal potency and causal resistance, including the correct order of rule emergence, temporary overgeneralization of an early rule, and increasing levels of potency and resistance. In the realm of seriation, the networks mimicked the development of four rule-based stages and a well-documented perceptual effect on differential stick sizes. Networks also predicted a new perceptual effect based on array disorder, for which some confirming evidence has been found. In the case of integration of time, distance, and velocity cues, the networks captured known rule-based stages and predicted reasonable new ones. In the one area of language development we simulated, acquisition of personal pronouns, network models simulated the beneficial effect of listening to overheard speech, as opposed to speech addressed to the child. This entailed a stage transition from persistent reversal errors to correct usage in the case of children exposed to a large proportion of directly addressed speech, and nearly errorless performance in the case of children exposed to a large proportion of overheard speech.

To appreciate the relations among the different simulation topics, it is useful to note that they can be classified in terms of whether the inputs and outputs are quantitative or qualitative in nature. All of the simulations require the integration of two or more distinct sources of input information to predict some result or action. The balance scale task requires integration of quantitative information on the weight and distance of objects on two sides of the scale to produce a qualitative prediction of which side of the scale will tip down. The causal prediction task requires integration of quantitative information on causal potency and causal resistance to quantitatively predict effect size. Seriation requires integration of quantitative information on the size and position of each of a number of sticks to yield a qualitative move: which stick to move and where to move it. The time-distance-velocity task requires integration of quantitative information on two dimensions to predict the quantitative value of a third dimension. Use of the personal pronouns *me* and *you* requires integration of qualitative information on the identities of speaker, addressee, and referent to yield a qualitative selection of the correct pronoun. This classification of tasks can be useful in interpreting perceptual effects. We now turn to a discussion of key features that can be abstracted from the various simulations.

Rule Learning

As noted in the review of previous work, connectionist learning has been demonstrated to generate rulelike behavior without explicit representation of rules. All of the present simulations show such rulelike behavior: the four balance scale rules; the two rules for predicting effect size from causal potency and resistance; the four seriation rules; the numerous normative and primitive rules for integrating time, distance, and velocity cues; and the semantic rules governing the use of English personal pronouns. In none of these cases were the relevant rules explicitly represented in the networks. Rather, rule use had to be diagnosed from systematic behavior just as is commonly done in psychological research with children. Explicit in the networks are units and their interconnections and activations. Patterns of network activity, after appropriate changes in network topology and weights, are capable of causing the network to behave as if it were following rules. Of course, the same might be true of children.

The basis for rulelike behavior in connectionist models is the ability of these networks to extract statistical regularities in the learning environment. In some cases, these regularities are simple linear relationships, but in the more interesting cases the regularities are subtle nonlinear relationships. In both cases, the network learns to similarly treat similar input-output pairs. But in the nonlinear cases, signaled in cascade-correlation nets by the recruitment of hidden units, the network must learn an underlying similarity structure that is not evident from examining the input patterns alone. Rather, the net's ability to represent abstract nonlinear similarity structures is developed through learning.

Unlike most previous models in cognitive development, the rulelike behavior of connectionist networks is learned, rather than hand designed by the modeler.⁵ Several previous models of the balance scale and seriation, for example, were characterized by hand-designed explicit rules that captured particular stages but failed to develop in the sense of moving naturally from one set of rules to the next.

Does rulelike behavior in connectionist networks mean that the networks are merely different implementations of symbolic rule-based systems (Fodor & Pylyshyn, 1988)? Not if the implementation differences make a difference in terms of ability to predict and interpret psychological phenomena. For example, one advantage of a connectionist implementation of rulelike behavior is that the "rules" are less brittle, in the sense of being more tolerant of input coding error, and more likely to generalize appro-

⁵A notable exception among explicit rule-based systems is Soar (Newell, 1990), a production system that learns its own rules through look-ahead search.

priately to novel patterns. Both minor deviations from trained inputs and novel, but similar, inputs would be handled in a mostly correct fashion by connectionist nets that have not simply memorized the training patterns. Simple memorization of training patterns can occur if the network has excess computational power (not likely in cascade-correlation nets because they recruit only the power they need) and is trained too deeply. Appropriate generalization to novelty was demonstrated in all of the present simulations. Other "implementational" advantages of connectionist nets are discussed later.

Stages

Not only do cascade-correlation nets learn appropriate rulelike behaviors, they can also learn them in psychologically realistic sequences. Such invariant sequences are a hallmark of stages of cognitive development (Flavell, 1971). Getting stages in the correct developmental order is not easy for computational modelers. For the domains treated here, even those previous models with a transition mechanism (all on the balance scale) failed to capture all of the stages in the correct order. Langley's (1987) rule-based model of the balance scale captured only Stage 3, ignoring Stages 1 and 2, and failing to reach Stage 4. The Soar model of the balance scale failed to progress past Stage 3 (Newell, 1990). The back-propagation model of the balance scale did reach Stage 4, but failed to stay there (McClelland, 1989). Reaching and staying in the final stage on the balance scale is partly a matter of learning the problem sufficiently deeply, but it is also a matter of being able to construct new representational power, something with which rule-based models and static connectionist networks have had considerable difficulty. Both the Langley model and the Soar model lacked the ability to represent torques and could not apparently develop this ability. The static back-propagation network model of the balance scale could not reach and stay in Stage 4 without sacrificing Stages 1 and 2.

Cascade-correlation nets captured the correct stage progressions in all five domains that were studied here, and did so with a variety of desirable psychological properties, including soft transitions, some stage skipping, and a limited amount of regression to early stages. Again, one notices less brittleness than is common in explicit rule-based models. With network models, transition to a higher level stage is typically soft and tentative, occasionally a stage is skipped altogether, and sometimes the net reverts back to an earlier stage at least temporarily. These tendencies reflect the dynamic, chaotic quality of connectionist networks. Due to the randomness inherent to starting configurations and other vagaries of travel through weight space and topology space, network behaviors are not entirely clear-cut and uniform.

It was argued elsewhere that connectionist models tend to produce those aspects of stages that are supported by psychological data (qualitative change, ordinality, and organization), and avoid producing those aspects of stages that are inconsistent with psychological data (abruptness and concurrence; Shultz, 1991). Networks do undergo qualitative changes in their behavior, capture invariant sequences of behavior, and exhibit a good deal of organization; but they change stages somewhat gradually and, because they are typically task specific, do not change stages all at once across domains.

The ability of cascade-correlation nets to capture correct stage sequences was due to a variety of factors. In the case of the balance scale, it was critical for the net to be in a particular region of connection weight space early in its developmental history and to recruit a small number of hidden units. The critical region of connection weight space is characterized by an emphasis on how much things weigh, as opposed to where they are placed on the balance scale, and could be managed either by environmental bias in favor of equal distance problems or by prestructuring network connections. With the current focus in the connectionist literature on determining what can be accomplished by learning from scratch, innate ideas are not much explored. But the potential for such investigation exists. Some connectionist researchers are beginning to study the breeding of connectionist networks through genetic algorithms (Belew et al., 1990) and the interaction of evolutionary processes with learning (Nolfi et al., 1990).

Stage sequences on the potency and resistance task and the time-distance-velocity task were the result of a network activation rule that sums its inputs. Hidden unit recruitment was also required for the achievement of higher level stages in the integration of time, distance, and velocity cues. Seriation stages resulted from a modularization of the task into selecting versus moving a stick and slight environmental biases in favor of smaller, less disordered arrays. The sequence of stages in the acquisition of personal pronouns (persistent reversal errors followed by correct usage) was due to an environmental bias for hearing directly addressed speech and the recruitment of a small number of hidden units.

Perceptual Effects

Connectionist techniques can also capture a variety of perceptual effects on cognitive developmental tasks. In the present work, these include the torque difference effect on the balance scale and the stick size effect in seriation. Such perceptual effects can be expected to occur whenever two or more items, quantitatively described, are being mapped onto a qualitative comparison. In contrast, no such effects can be expected on tasks like potency and resistance or time-distance-velocity, where quantitative inputs

predict a quantitative output. No matter how distinctive the input values are for these quantitative to quantitative tasks, the idea is to predict the output value as precisely as possible.

Such perceptual effects are pervasive in cognitive developmental research, but no theoretical account integrates them with the cognitive features of the task. Perceptual effects appear to be particularly immune to symbolic rule-based accounts because rules are typically sensitive only to the direction of input differences, not to the amount of such differences. For example, a balance scale rule might be sensitive to whether one side of the scale had more weights than the other side, but it would not typically be sensitive to how much more. In contrast, the naturalness of the emergence of these perceptual effects from connectionist models is worth noting. Perceptual effects are a natural result of the continuous nature of network computations. Larger differences in inputs produce clearer activation patterns on hidden units and more decisive qualitative decisions on output units. Connectionist accounts hold the promise of a much tighter theoretical integration of perceptual and cognitive factors than was previously possible. This is another case of where connectionist implementations have a decided implementational and explanatory advantage over rule-based implementations.

Theoretical Issues

Connectionist models are not psychological theories. Rather, they are powerful tools that may, along with more conventional empirical and theoretical work, help us to develop coherent psychological theories (McCloskey, 1991). Good theories explain phenomena in terms of independently motivated principles and show how previously unrelated phenomena derive from common underlying principles. Theoretical ideas emanating from successful connectionist models often satisfy these criteria (Seidenberg, 1993). Here, the explanatory principles are constraints on cognitive development—some, domain-general and others, domain-specific.

Among the domain-general constraints: (a) Cognitive judgments, decisions, and actions result from brain-style computation, in which excitation or inhibition is passed among simple processing units that vary in their temporary level of activity; and (b) Cognitive developmental transitions occur through the dual techniques of connection weight adjustments between existing units and recruitment of new hidden units, both of which serve to reduce the discrepancy between expectations and results. In the present models, these domain-general constraints are formally specified by the cascade-correlation algorithm.

Among the domain-specific constraints: (a) Some problems are too difficult to solve in a single homogenous network. Such problems require

modularity in network organization, such that different networks solve different aspects of the overall problem. (b) Different problems require different coding schemes for inputs and outputs in the training and generalization patterns. Although many phenomena appear to be robust against considerable variation in coding techniques, different problems do require particular input data and output actions. (c) Some phenomena require bias in either the training environment or the initial weight structure.

These domain-specific constraints can be taken as testable predictions whenever they are not yet empirically documented. Such constraints are similar to Gelman's (1990) *first principles* that, to enable learning, focus the child's attention on the relevant features of the environment. This type of constraint does not force the child to attend to particular data, but it does provide a filter for data relevance.

We differ from the more extreme nativist positions in the belief that our particular input representations are not necessarily innate. We do not claim that children are born with knowledge representations innately adapted for tasks such as seriation or balance scales, but rather, their performance on these tasks results from a process that incorporates these types of input constraints. That is, when children become able to learn about tasks like seriation or the balance scale, they do so under the kinds of input constraints described by our models. We leave open the question of whether these specified input configurations result from a maturational process or from earlier learning (e.g., during infancy) that itself might have been highly constrained.

Weight adjustment is suited to modeling underlying quantitative changes, whereas hidden unit recruitment is suited to modeling qualitative ones. This allows a novel and computationally precise reformulation of Piaget's useful, but vague, notions of assimilation and accommodation. Indeed, we can now go one step farther than did Piaget, because he had no way of describing learning without accommodation, that is, without qualitative change.

Using Piaget's terms, one can conceptualize three general types of cognitive encounters in cascade-correlation nets: assimilation, assimilative learning, and accommodation. Pure assimilation occurs without learning. It is represented in cascade-correlation by correct generalization to novel problems without either weight changes or hidden unit recruitment. Assimilative learning occurs by weight adjustment, but without hidden unit recruitment. Here, the network learns new patterns that do not require nonlinear changes in representational power. Accommodation occurs via hidden unit recruitment when new patterns cannot be learned without nonlinear increases in computational power.

In cascade-correlation, these three types of encounter are all driven by the

same process—adaptation via error reduction. Although these three possibilities can be conceptualized as qualitatively different from each other, it is perhaps more useful to view them quantitatively, on a dimension of learning difficulty. Pure assimilation is easiest because it requires little or no learning, whereas accommodation is relatively difficult, as assessed by metrics such as epochs to learn.

Adaptation through assimilation and accommodation can also be reinterpreted through rule-based and back-propagation perspectives, but with less satisfactory results. In a rule-based learning system like Soar, assimilation could be construed as rule-firing, and accommodation, as chunking new rules through impasse-driven search (Newell, 1990). In back-propagation learning, accommodation could be viewed in terms of weight adjustment and assimilation as the absence of such adjustment (McClelland, 1989). There is room for assimilative learning in neither of these frameworks. All learning in rule-based systems seems to create qualitatively different structures. Each new rule adds qualitatively different structure (van Geert, 1991). Conversely, no learning in static back-propagation nets creates qualitatively different structures because the network topology never changes. Yet, psychologically, some learning requires small quantitative adjustments (e.g., learning a new phone number), whereas other learning requires more substantial qualitative changes in representation and processing (e.g., learning to use personal pronouns or learning to integrate time, distance, and velocity information).⁶

Thus, on theoretical grounds, cascade-correlation is a particularly promising tool for modeling and eventually explaining cognitive development. The simulations reported here suggest that cascade-correlation networks can capture a wide range of developmental phenomena. Such successful applications could lead to new explanatory theories of cognitive development, although these new theories might look rather different than do classical information processing accounts (Seidenberg, 1993).

In broad outline, such a connectionist-inspired theory views the child as being equipped with powerful, general purpose learning techniques, based primarily on pattern association, but capable of constructing new represen-

⁶Klahr's chapter 8 (this volume) suggests that computational analysis of Piaget's notions of assimilation and accommodation is uninformative. Our motivation for this analysis is to show that the mathematics of cascade-correlation can be related to some traditional, albeit vague ideas about transition in cognitive development. This might enhance the relevance of our simulations for traditional developmentalists, but more significantly it underscores the importance of having a transition rule that incorporates both qualitative and quantitative change. Also, the mapping of computational models to assimilation and accommodation is not as unconstrained as Klahr suggests. The developmental models presented throughout this volume are each sufficiently specified to determine whether they incorporate qualitative or quantitative change.

tations and thus, greater computational power. Knowledge is represented by patterns of activation across many simple processing units, not by explicitly formulated symbolic rules. Cognitive processing occurs according to basic principles of neuronal functioning (excitation and inhibition) rather than by the matching and firing of rules. Rather than an artificial separation between perceptual and cognitive processes, there is a tight theoretical integration of perception and cognition. Such a system can be innately structured in certain ways and learns from environmental feedback, based primarily on correlations among events and instances, with sensitivity to biases afforded by the learning environment. Qualitatively new representational skills emerge as required to reduce error. When new representational structures do emerge, they elaborate on the results of earlier computations. Learning is the primary engine of cognitive transitions, yielding the many qualitative and quantitative changes one sees in cognitive development.

ACKNOWLEDGMENTS

This work was supported by grants from the Natural Science and Engineering Research Council of Canada and the Fonds pour la Formation de Chercheurs et l'Aide à la Recherche du Québec.

REFERENCES

- Alpaydin, E. (1991). *GAL: Networks that grow when they learn and shrink when they forget* (Tech. Rep. No. 91-032). Berkeley, CA: International Computer Science Institute.
- Anderson, N. H. (1974). Information integration theory: A brief survey. In D. H. Krantz, R. C. Atkinson, R. D. Luce, & P. Suppes (Eds.), *Contemporary developments in mathematical psychology* (Vol. 2, pp. 236-305). San Francisco: Freeman.
- Anderson, N. H. (1991). Functional memory in person cognition. In N. H. Anderson (Ed.), *Contributions to information integration theory: Vol. 1. Cognition* (pp. 1-55). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Bates, E. A., & Elman, J. L. (1993). Connectionism and the study of change. In M. H. Johnson (Ed.), *Brain development and cognition* (pp. 623-642). Oxford, England: Blackwell.
- Baylor, G. W., Gascon, J., Lemoyne G., & Pothier N. (1973). An information-processing model of some seriation tasks. *Canadian Psychologist*, 14, 167-196.
- Belew, R., McInery, J., & Schraudolph, N. N. (1990). *Evolving networks: Using the genetic algorithm with connectionist learning* (Tech. Rep. No. CS90-174). San Diego: University of California, Cognitive Computer Science Research Group, Computer Science and Engineering Department.
- Bettleheim, B. (1967). *The empty fortress: Infantile autism and the birth of the self*. New York: The Free Press.
- Boden, M. A. (1980). Artificial intelligence and Piagetian theory. In M. Boden (Ed.), *Minds*

- and mechanisms: *Philosophical psychology and computational models* (pp. 236-261). Ithaca, NY: Cornell University Press.
- Boden, M. A. (1982). Is equilibration important? A view from artificial intelligence. *British Journal of Psychology*, 73, 65-173.
- Boden, M. A. (1988). *Computer models of mind*. New York: Cambridge University Press.
- Buckingham, D., & Shultz, T. R. (1994). A connectionist model of the development of velocity, time, and distance concepts. In *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society* (pp. 72-77). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Charney, R. (1980a). Pronoun errors in Autistic children: Support for a social explanation. *British Journal of Disorders of Communication*, 15, 39-43.
- Charney, R. (1980b). Speech roles and the development of personal pronouns. *Journal of Child Language*, 7, 509-528.
- 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.
- Chiat, S. (1981). Context-specificity and generalization in the acquisition of pronominal distinctions. *Journal of Child Language*, 8, 75-91.
- Chletsos, P. N., De Lisi, R., Turner, G., & McGillicuddy-De Lisi, A. V. (1989). Cognitive assessment of proportional reasoning strategies. *Journal of Research and Development in Education*, 22, 18-27.
- Clark, E. V. (1978). From gesture to word: On the natural history of deixis in language acquisition. In J. S. Bruner & A. Garton (Eds.), *Human growth and development* (pp. 85-120). Oxford, England: Oxford University Press.
- de Paulo, B. M., & Bonvillian, J. D. (1978). The effect on language development of the special characteristics of speech addressed to children. *Journal of Psycholinguistic Research*, 7, 189-211.
- Dudai, Y. (1989). *The neurobiology of memory: Concepts, findings, and trends*. Oxford, England: Oxford University Press.
- Elkind, D. (1964). Discrimination, seriation, and numeration of size and dimensional differences in young children: Piaget replication study VI. *Journal of Genetic Psychology*, 104, 276-296.
- Elman, J. (1993). Learning and development in neural networks: The importance of starting small. *Cognition*, 48, 71-99.
- Ervin-Tripp, S. (1971). An overview of theories of grammatical development. In D. I. Slobin (Ed.), *The ontogenesis of grammar: A theoretical symposium* (pp. 189-212). New York: Academic Press.
- Fahlman, S. E. (1988). Faster-learning variations on back-propagation: An empirical study. In D. S. Touretzky, G. E. Hinton, and T. J. Sejnowski (Eds.), *Proceedings of the 1988 Connectionist Models Summer School* (pp. 38-51). Los Altos, CA: Morgan Kaufmann.
- Fahlman, S. E., & Lebiere C. (1990). The cascade-correlation learning architecture. In D. S. Touretzky (Ed.), *Advances in neural information processing systems 2* (pp. 524-532). Los Altos, CA: Morgan Kaufmann.
- Ferretti, R. P., & Butterfield, E. C. (1986). Are children's rule assessment classifications invariant across instances of problem types? *Child Development*, 57, 1419-1428.
- Flavell, J. H. (1971). Stage-related properties of cognitive development. *Cognitive Psychology*, 2, 421-453.
- Flavell, J. H. (1984). Discussion. In R. J. Sternberg (Ed.), *Mechanisms of cognitive development* (pp. 187-209). New York: Freeman.
- Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28, 3-71.
- Gelman, R. (1990). First principles organize attention to and learning about relevant data: Number and the animate-inanimate distinction as examples. *Cognitive Science*, 14, 79-106.

- Hare, M., & Elman, J. L. (1992). A connectionist account of English inflectional morphology: Evidence from language change. In *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society* (pp. 265-270). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Harnad, S. Hanson, S.J., & Lubin, J. (1994). Learned categorical perception in neural nets: Implications for symbol grounding. In V. Honavar & L. Uhr (Eds.), *Artificial intelligence and neural networks: Steps toward principled integration* (pp. 191-206). New York: Academic Press.
- Hertz, J., Krogh, A., & Palmer, R. G. (1991). *Introduction to the theory of neural computation*. New York: Addison-Wesley.
- Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence*. New York: Basic Books.
- Jacobs, R. A., Jordan, M. I., & Barto, A. G. (1991). Task decomposition through competition in a modular connectionist architecture: The what and where vision tasks. *Cognitive Science*, 15, 219-250.
- Kingma, J. (1982). A criterion problem: The use of different operationalizations in seriation research. *Perceptual and Motor Skills*, 55, 1303-1316.
- Kingma, J. (1984). The influence of task variations in seriation research: Adding irrelevant cues to the stimulus materials. *Journal of Genetic Psychology*, 144, 241-253.
- Klahr, D., Langley, P., & Neches R. (Eds.). (1987). *Production system models of learning and development*. Cambridge, MA: MIT Press.
- Klahr, D., & Siegler, R. S. (1978). The representation of children's knowledge. In H. W. Reese & L. P. Lipsitt (Eds.), *Advances in child development and behavior* (pp. 61-116). New York: Academic Press.
- Koslowski, B. (1980). Quantitative and qualitative changes in the development of seriation. *Merrill-Palmer Quarterly*, 26, 391-405.
- Langley, P. (1987). A general theory of discrimination learning. In D. Klahr, P. Langley, & R. Neches (Eds.), *Production system models of learning and development* (pp. 99-161). Cambridge, MA: MIT Press.
- Levin, I. (1977). The development of time concepts in young children: Reasoning about duration. *Child Development*, 48, 435-444.
- Levin, I. (1979). Interference of time-related and unrelated cues with duration comparisons of young children: Analysis of Piaget's formulation of the relation of time and speed. *Child Development*, 50, 469-477.
- Lewandowsky, S. (1993). The rewards and hazards of computer simulations. *Psychological Science*, 4, 236-243.
- Macnamara, J. (1976). Stomachs assimilate and accommodate, don't they? *Canadian Psychological Review*, 3, 67-173.
- MacWhinney, B., Leinbach, J., Taraban, R., & McDonald, J. (1989). Language learning: Cues or rules? *Journal of Memory and Language*, 28, 255-277.
- Marchman, V. A. (1992). *Language learning in children and neural networks: Plasticity, capacity, and the critical period* (Tech. Rep. No. 9201). San Diego: University of California, Center for Research in Language.
- Mareschal, D. (1991). *Cascade-correlation and the genetron: Possible implementations of equilibration* (Tech. Rep. 91-10-17). Montréal: McGill University, McGill Cognitive Science Centre.
- Mareschal, D. (1992). *A connectionist model of the development of children's seriation abilities*. Unpublished master's thesis. McGill University, Montréal, Canada.
- Mareschal, D., & Shultz, T. R. (1993). A connectionist model of the development of seriation. In *Proceedings of the Fifteenth 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. G. M. Morris (Ed.), *Parallel distributed processing: Implications for*

- psychology and neurobiology* (pp. 8-45). Oxford, England: Oxford University Press.
- McClelland, J. L., & Jenkins, E. (1991). Nature, nurture, and connections: Implications of connectionist models for cognitive development. In K. VanLehn (Ed.), *Architectures for intelligence* (pp. 41-73). Hillsdale, NJ: Lawrence Erlbaum Associates.
- McCloskey, M. (1991). Networks and theories: The place of connectionism in cognitive science. *Psychological Science*, 2, 387-395.
- Minsky, M. (1986). *The society of mind*. New York: Simon & Schuster.
- Minsky, M., & Papert, S. (1969). *Perceptrons*. Cambridge, MA: MIT Press.
- Newell, A. (1990). *Unified theories of cognition*. Cambridge, MA: Harvard University Press.
- Ninio, A. (1979). Piaget's theory of space perception in infancy. *Cognition*, 7, 125-144.
- Nolfi, S., Elman, J. L., & Parisi, D. (1990). *Learning and evolution in neural networks* (Tech. Rep. No. 9019). San Diego: University of California, Center for Research in Language.
- Oden, G. C. (1987). Concept, knowledge, and thought. *Annual Review of Psychology*, 38, 203-227.
- Oshima-Takane, Y. (1988). Children learn from speech not addressed to them: The case of personal pronouns. *Journal of Child Language*, 15, 95-108.
- Oshima-Takane, Y. (1992). Analysis of pronominal errors: A case study. *Journal of Child Language*, 19, 111-131.
- Oshima-Takane, Y., & Derevensky, J. L. (1990, April). *Do later-born children delay in early language development?* Paper presented at the International Conference on Infant Studies, Montréal, Canada.
- Papert, S. (1963). *Intelligence chez l'enfant et chez le robot*. [Intelligence in the child and the robot.] In L. Apostel, J. Grize, S. Papert, & J. Piaget. *La filiation des structures. Etudes D'Epistemologie Genetique*, 15, 131-194.
- Piaget, J. (1965). *The child's concept of number*. New York: Norton.
- Piaget, J. (1969). *The child's conception of time*. London: Routledge & Kegan Paul.
- Piaget, J. (1970). *The child's conception of movement and speed*. London: Routledge & Kegan Paul.
- Piaget, J. (1972). *Problèmes de psychologie génétique*. [Problems of genetic psychology.] Paris: Denoel/Gontier.
- Piaget, J. (1977). *The development of thought: Equilibration of cognitive structures*. Oxford, England: Blackwell.
- Piaget, J., & Inhelder, B. (1969). *The psychology of the child*. New York: Basic Books.
- Piaget, J., & Inhelder, B. (1973). *Memory and intelligence*. London: Routledge & Kegan Paul.
- Plunkett, K., & Marchman, V. (1991). U-shaped learning and frequency effects in a multi-layered perceptron: Implications for child language acquisition. *Cognition*, 38, 43-102.
- Plunkett, K., & Sinha, C. (1992). Connectionism and developmental theory. *British Journal of Developmental Psychology*, 10, 209-254.
- Pollack, J. B. (1990). Language acquisition via strange automata. In *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society* (pp. 678-685). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Retschitzki, J. (1978). L'évolution des procédures de sériation: Étude génétique et simulation. [Evolution of seriation procedures: Developmental study and simulation.] *Archives de Psychologie*, 46, Monographie 5.
- Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 1, pp. 318-362). Cambridge, MA: MIT Press.
- Schiff-Meyers, N. (1983). From pronoun reversals to correct pronoun usage: A case study of a normally developing child. *Journal of Speech and Hearing Disorders*, 48, 385-394.

- Schmidt, W. C. (1991). Connectionist models of balance scale phenomena. Unpublished honours thesis, McGill University, Montréal, Canada.
- Schmidt, W. C., & Shultz, T. R. (1991). *A replication and extension of McClelland's balance scale model* (Tech. Rep. No. 91-10-18). Montréal: McGill University, McGill Cognitive Science Centre.
- Schyns, P. (1991). A modular neural network model of concept acquisition. *Cognitive Science*, 15, 461-508.
- Seidenberg, M. S. (1993). Connectionist models and cognitive theory. *Psychological Science*, 4, 228-235.
- Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review*, 96, 523-568.
- Shipley, E. F., & Shipley, T. E. (1969). Quaker children's use of *Thee*: A relational analysis. *Journal of Verbal Learning and Verbal Behavior*, 8, 112-117.
- Shultz, T. R. (1991). Simulating stages of human cognitive development with connectionist models. In L. Birnbaum & G. Collins (Eds.), *Machine learning: Proceedings of the Eighth International Workshop* (pp. 105-109). San Mateo, CA: Morgan Kaufman.
- Shultz, T. R., Buckingham, D., & Oshima-Takane, Y. (1994). A connectionist model of the learning of personal pronouns in English. In S. J. Hanson, T. Petsche, M. Kearns, & R. L. Rivest (Eds.), *Computational learning theory and natural learning systems: Intersection between theory and experiment* (Vol. 2, pp. 347-362). Cambridge, MA: MIT Press.
- Shultz, T. R., Mareschal, D., & Schmidt, W. C. (1994). Modeling cognitive development on balance scale phenomena. *Machine Learning*, 16, 57-86.
- Shultz, T. R. & Schmidt, W. C. (1991). A cascade-correlation model of balance scale phenomena. In *Proceedings of the Thirteenth Annual Conference of the Cognitive Science Society* (pp. 635-640). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Shultz, T. R., Zelazo, P. R., & Strigler, D. (1991, April). *Connectionist modeling of the development of the concepts of potency and resistance in causal prediction*. Paper presented at the meeting of the Society for Research in Child Development, Seattle, WA.
- Siegler, R. S. (1976). Three aspects of cognitive development. *Cognitive Psychology*, 8, 481-520.
- Siegler, R. S. (1981). Developmental sequences between and within concepts. *Monographs of the Society for Research in Child Development*, 46 (Whole No. 189).
- Siegler, R. S., & Richards, D. D. (1979). Development of time, speed, and distance concepts. *Developmental Psychology*, 15, 288-298.
- Spelke, E. (1990). Principles of object perception. *Cognitive Science*, 14, 29-56.
- Squire, L. (1987). *Memory and brain*. Oxford, England: Oxford University Press.
- van der Maas, H. L. J., & Molenaar, P. C. M. (1992). Stagewise cognitive development: An application of catastrophe theory. *Psychological Review*, 99, 395-417.
- van Geert, P. (1991). A dynamic systems model of cognitive and language growth. *Psychological Review*, 98, 3-53.
- Wilkening, F. (1981). Integrating velocity, time, and distance information: A developmental study. *Cognitive Psychology*, 13, 231-247.
- Wilkening, F. (1982). Children's knowledge about time, distance, and velocity interrelations. In W. J. Friedman (Ed.), *The developmental psychology of time* (pp. 87-112). New York: Academic Press.
- Young, R. (1976). *Seriation by children: An artificial intelligence analysis of a Piagetian task*. Basel, Switzerland: Birkhauser.
- Zelazo, P. D., & Shultz, T. R. (1989). Concepts of potency and resistance in causal prediction. *Child Development*, 60, 1307-1315.

