Brain and Mind **4:** 129–149, 2003.

Interpreting the Internal Structure of a Connectionist Model of the Balance Scale Task

MICHAEL R. W. DAWSON[1] and CORINNE ZIMMERMAN[2]
[1]*Biological Computation Project, University of Alberta, Edmonton, Alberta, Canada T6G 2E9, e-mail: mdawson@ualberta.ca, www.bcp.psych.ualberta.ca/∼mike/*
[2]*Department of Psychology, Illinois State University, Campus Box 4620, Normal, IL 61790-4620, U.S.A., e-mail: czimmer@ilstu.edu*

(Received: 9 March, 2003; in final form: 23 April, 2003)

Abstract. One new tradition that has emerged from early research on autonomous robots is embodied cognitive science. This paper describes the relationship between embodied cognitive science and a related tradition, synthetic psychology. It is argued that while both are synthetic, embodied cognitive science is antirepresentational while synthetic psychology still appeals to representations. It is further argued that modern connectionism offers a medium for conducting synthetic psychology, provided that researchers analyze the internal representations that their networks develop. The paper then provides a detailed example of the synthetic approach by showing how the construction (and subsequent analysis) of a connectionist network can be used to contribute to a theory of how humans solve Piaget's classic balance scale task.

Key words: balance scale task, cognitive informatics, connectionism, embodied cognitive science, synthetic psychology.

1. Embodied Cognitive Science

Cognitive informatics is an interdisciplinary study of cognition, perception, and action. It is based on the assumption that cognition is information processing (Dawson, 1998), where information processing is generally construed as the rule-governed manipulation of data structures that are stored in a memory.

Of course, not all researchers are comfortable with adopting this research program, because they have fundamental disagreements with this foundational assumption. For example, the embodied cognitive science movement challenges the symbol-based conception of cognitive processing by using many of the same arguments that were employed by connectionist researchers in the early 1980s. Embodied cognitive science is a reaction against the traditional view that human beings as information processing systems "receive input from the environment (perception), process that information (thinking), and act upon the decision reached (behavior). This corresponds to the so-called sense-think-act cycle" (Pfeifer and Scheier, 1999). The *sense–think–act cycle*, which is a fundamental characteristic of conventional theories of cognition, is an assumption that the embodied approach considers to be fatally flawed.

Embodied cognitive science argues that theories of intelligence should exhibit two basic characteristics. First, they should be embodied, meaning that the theory should take the form of a working computer simulation or robot. Second, they should be situated, meaning that the simulation or robot should have the capability of sensing its environment. Embodied cognitive scientists create embodied, situated agents in order to create novel and surprising behaviors that emerge from the interaction between agents and their environments. One of the aims of embodied cognitive science is to replace the sense–think–act cycle with mechanisms of sensory-motor coordination (Pfeifer and Scheier, 1999) that might be construed as forming a *sense–act cycle*. The purpose of this change is to reduce, as much as possible, the role of internal representations in mediating intelligence. If one situates an autonomous agent in such a way that the agent can sense the world, then no internal representation of the world is necessary. "The realization was that the so-called central systems of intelligence — or core AI as it has been referred to more recently — was perhaps an unnecessary illusion, and that all the power of intelligence arose from the coupling of perception and actuation systems" (Brooks, 1999).

1.1. HISTORICAL EXAMPLES OF THE EMBODIED APPROACH

One reason that embodied cognitive science is attractive and is growing in popularity is because it can call on a long history of success stories in which extremely interesting behaviors emerged from fairly simple devices.

1.1.1. *The Homeostat*

One important historical example of the embodied approach comes from Ashby's study of feedback between generic machines (Ashby, 1960). For Ashby, a machine was simply a device which, when given a particular input, generates a corresponding output. Of particular interest to Ashby was a system of four different machines coupled together with feedback. Ashby realized that this system was sufficiently complex that it could not be studied analytically. Ashby (1960) dealt with this problem by constructing a device, called the homeostat, which allowed him to observe the behavior of this complicated set of feedback relationships. In other words, he adopted a synthetic approach for exploring feedback of this type.

In general the homeostat was a device that monitored its own internal stability (i.e., the amount of current being generated by each of its four component devices). If subjected to external forces, such as an experimenter manipulating one of its four component machines by hand, this internal stability was disrupted and the homeostat was moved into a higher energy, less stable state. When this happened, the homeostat would modify the internal connections between its component units by advancing one or more of its internal switches to modify the states of its internal potentiometers, which essentially served as 'connection weights' for the signals being sent between its component machines. The modified potentiometer settings enabled the homeostat to return to a low energy, stable state. The homeostat was "like a fireside cat or dog which only stirs when disturbed, and then methodically finds a comfortable position and goes to sleep again" (Grey Walter, 1963).

The homeostat was tested by placing some of its components under the direct control of the experimenter, by manipulating these components, and by observing the changes in the system as a whole. Even with this fairly simple pattern of feedback amongst four component devices, many surprising emergent behaviors were observed. For example, in one interesting study Ashby (1960) demonstrated that the system was capable of a simple kind of learning. Ashby went on to demonstrate that the homeostat was also capable of adapting to two different environments that were alternated.

1.1.2. *The Tortoise*

Ashby's homeostat could be interpreted as supporting the claim that the complexity of the behavior of whole organisms largely emerges from (a) a large number of internal components and from (b) the interactions between these components. In the late 1940s, some of the first autonomous robots — called tortoises because of their appearance — were built to investigate a counterclaim (Grey Walter, 1950, 1951, 1963). Grey Walter's research program "held promise of demonstrating, or at least testing the validity of, the theory that multiplicity of units is not so much responsible for the elaboration of cerebral functions, as the richness of their interconnection" (Grey Walter, 1963). His goal was to use a very small number of components to create robots that generated much more life-like behavior than that exhibited by Ashby's homeostat.

At a general level, a tortoise was an autonomous motorized tricycle. One motor was used to rotate the two rear wheels forward. The other motor was used to steer the front wheel. The behavior of these two motors was under the control of two different sensing devices. The first was a photoelectric cell that was mounted on the front of the steering column, and which always pointed in the direction that the front wheel pointed. The other was an electrical contact that served as a touch sensor. This contact was closed whenever the transparent shell that surrounded the rest of the robot encountered an obstacle.

Of a tortoise's two reflexes, the light-sensitive one was the more complex. In low light, the rear motor would propel the robot forward while the steering motor slowly turned the front wheel. As a result, the machine could be described as exploring its environment. When moderate light was detected by the photoelectric cell, the steering motor stopped. As a result, the robot moved forward, approaching the source of the light. However, if the light source were too bright, then the steering motor would be turned on again at twice the speed that was used during the robot's exploration of the environment. As a result, "the creature abruptly sheers away and seeks a more gentle climate. If there is a single light source, the machine circles around it in a complex path of advance and withdrawal" (Grey Walter, 1950).

The touch reflex that was built into a tortoise was wired up in such a way that when it was activated, any signal from the photoelectric cell was ignored. When the tortoise's shell encountered an obstacle, an oscillating signal was generated that rhythmically caused both motors to run at full power, turn off, and to run at full power again. As a result, "all stimuli are ignored and its gait is transformed into a succession of butts, withdrawals and sidesteps until the interference is either pushed aside or circumvented" (Grey Walter, 1950).

In spite of their simple design, Grey Walter demonstrated that his robots were very capable of complex and interesting behaviors. One of his tortoises could move around an obstacle, and then orbit a light source with complicated movements that would not take it too close, but also would not take it too far away. If presented two light sources, complex choice behavior was observed. If a tortoise encountered a mirror, then a light source mounted on top of the robot became a stimulus for its light sensor, and resulted in what became known as the famous mirror dance, in which the robot "lingers before a mirror, flickering, twittering and jigging like a clumsy Narcissus. The behavior of a creature thus engaged with its own reflection is quite specific, and on a purely empirical basis, if it were observed in an animal, might be accepted as evidence of some degree of self-awareness" (Grey Walter, 1963).

1.2. THE SYNTHETIC APPROACH

Most models in classical cognitive science and in experimental psychology are derived from the analysis of existing behavioral measurements (Dawson, 2003). In contrast, both the homeostat and the tortoise are examples of a much more synthetic approach to research. They both involved making some assumptions about primitive capacities, which were then built into working systems whose behaviors were observed. In the synthetic approach, model construction *precedes* behavioral analysis.

Braitenberg (1984) has argued that psychology should adopt the synthetic approach, because theories that are derived via analysis are inevitably more complicated than is necessary. This is because cognitive scientists and psychologists have a strong tendency to ignore the parable of the ant, and prefer to locate the source of complicated behavior within the organism, and not within its environment. Pfeifer and Scheier (1999) call this the frame-of-reference problem. A consequence of the frame-of-reference problem is that because of nonlinear interactions (such as feedback between components, and between a system and its environment), relatively simple systems can surprise us, and generate far more complicated behavior than we might expect. The further appeal of the synthetic approach comes from the belief that if we have constructed this simple system, we should be in a very good position to propose a simpler explanation of its complicated behavior. In particular, we should be in a better position than would be the case if we started with the behavior, and attempted to analyze it in order to understand the workings of an agent's internal mechanisms.

Clearly, the synthetic approach is worth exploring, particularly if it offers the opportunity to produce simple theories of complex, and emergent, behaviors. For this reason, Braitenberg (1984) has called for the development of a new approach that he has named *synthetic psychology*. However, the synthetic approach as it appears in embodied cognitive science is associated with a view that many psychologists would not be comfortable in endorsing.

1.3. REACTING AGAINST REPRESENTATION

Modern embodied cognitive science can be viewed as a natural evolution of the historical examples that were presented earlier. Researchers have used the synthetic approach to develop systems that generate fascinatingly complicated behaviors (Braitenberg, 1984; Brooks, 1999; Pfeifer and Scheier, 1999). However, much of this research is dramatically antirepresentational. "In particular I have advocated situatedness, embodiment, and highly reactive architectures with no reasoning systems, no manipulable representations, no symbols, and totally decentralized computation" (Brooks, 1999). One of the foundational assumptions of behavior-based robotics is that if a system can sense its environment, then it should be unnecessary for the system to build an internal model of the world.

This is strongly reminiscent of a failed tradition in experimental psychology, called *behaviorism,* that attempted to limit psychological theory to observables (namely, stimuli and responses), and which viewed as unscientific any theories that attempted to describe internal processes that mediated relationships between sensations and actions (Watson, 1913). The resemblance of embodied cognitive science to behaviorism is unfortunate, because it decreases the likelihood that the advantages of the synthetic approach will be explored in psychology. The reason for this is that many higher order psychological phenomena require an appeal to internal representations in order to be explained.

2. Synthetic Psychology

2.1. THE NEED FOR REPRESENTATION

That stimulus–response reflexes are not sufficient to account for many higher order psychological phenomena is a theme that has dominated cognitivism's replacement of behaviorism as the dominant theoretical trend in experimental psychology. In the study of language, this theme was central to Chomsky's critical review (Chomsky, 1959) of Skinner (1957). Many of the modern advances in linguistics were the direct result of Chomsky's proposal that generative grammars provided the representational machinery that mediated regularities in language (Chomsky, 1965; Chomsky and Halle, 1991; Chomsky, 1995). Similar arguments were made against purely associationist models of memory and thought (Anderson and Bower, 1973). For example, Bever *et al.* (1968) formalized associationism as a finite state automaton, and demonstrated that such a system was unable to deal with the clausal structure that typifies much of human thought and language. Paivio (1969, 1971) used the experimental methodologies of the verbal learners to demonstrate that a representational construct — the imageability of concepts — was an enormously powerful predictor of human memory. The famous critique of 'old connectionism' by Minsky and Papert (1988) could be considered a proof about the limitations of visual systems that do not include mediating representations. These examples, and many more, have lead to the status quo view that representations are fundamental to cognition and perception (Fodor, 1975; Marr, 1982; Pylyshyn, 1984; Jackendoff, 1992; Dawson, 1998).

Some robotics researchers also share this sentiment, although it must be remembered that behavior-based robotics was a reaction against their representational work (Brooks, 1999). Moravec (1999) suggests that the type of situatedness that characterizes behavior-based robotics (for example, the simple reflexes that guided Grey Walter's tortoises) probably provides an accurate account of insect intelligence. However, at some point systems built from such components will have at best limited abilities. "Real insects illustrate the problem. The vast majority fails to complete their life cycles, often doomed, like moths trapped by a streetlight, by severe cognitive limitations." Internal representations are one obvious medium for surpassing such limitations. The question that this leads to is this: can the synthetic approach be conducted in a way that provides the advantages that have been raised above, but that also provides insight into representational processing?

2.2. CONNECTIONISM, SYNTHESIS, REPRESENTATION

Of course, the answer to this question is a resounding yes. There is nothing in the synthetic approach per se that prevents one from constructing systems that use representations. Describing a model as being synthetic or analytic is using a dimension that it is completely orthogonal to the dimension used when describing a model as being representational or not. Dawson (2003, Chapter 8) has provided a detailed argument that for the study of higher order cognition, researchers should adopt an approach that is both synthetic and representational. He then goes on to suggest that connectionist or parallel distributed processing (PDP) models are an attractive medium for carrying out this kind of research program.

2.2.1. *Connectionism in Brief*

This section provides a brief overview of the general properties of connectionist models. For a more detailed overview of these models, the reader is referred to Dawson (1998, Chapter 3). Dawson (2003, Chapters 9–11) provides a complete introduction to a variety of connectionist architectures.

A connectionist or PDP network is a system of interconnected, simple processing units that can be used to classify patterns presented to it. A PDP network is usually made up of three kinds of processing units: *Input units* encode the stimulus or activity pattern that the network will eventually classify; *hidden units* detect features or regularities in the input patterns, which can be used to mediate classification; and *output units* represent the network’s response to the input pattern (i.e., the category to which the pattern is to be assigned) on the basis of features or regularities that have been detected by the hidden units. Processing units communicate by means of sending signals through weighted connections.

In most cases, a processing unit carries out three central functions: First, a processor computes the total signal that it receives from other units. A *net input function* is used to carry out this calculation. After the processing unit determines its net input, it transforms it into an internal level of activity, which typically ranges between 0 and 1. The internal activity level is calculated by means of an *activation function*. Finally, the processing unit uses an *output function* to convert its internal activity into a signal to be sent to other units.

The signal sent by one processor to another is transmitted through a weighted connection, which is typically described as being analogous to a synapse. The connection itself is merely a communication channel. The weight associated with the connection defines its nature and strength. For example, inhibitory connections are defined with negative weights, and excitatory connections are defined with positive weights. A strong connection has a weight with a large absolute value, while a weak connection has a weight with a near-zero absolute value. The pattern of connections in a PDP network defines the clausal relations between the processors and is therefore analogous to a program in a conventional computer (Smolensky, 1988).

Unlike a conventional computer, though, a network is not given a step-by-step procedure for performing a desired task. It is instead *trained* to solve the task on its own. For instance, consider a popular supervised learning procedure called the *generalized delta rule* (Rumelhart *et al.*, 1986). To train a network with the generalized delta rule, one begins with a network that has small, randomly assigned connection weights. The network is then presented a set of training patterns, each of which is paired with a known desired response. To train a network on one of these patterns, the pattern is presented to the network's input units, and the network generates a response using its existing connection weights. An error value for each output unit is then calculated by comparing the actual output to the desired output. This error value is then used to modify connection weights in such a way that the next time this pattern is presented to the network, the network's output errors will be smaller. By repeating this procedure a large number of times for each pattern in the training set, the network's response errors for each pattern can be reduced to near zero. At the end of this procedure, the network will have a very specific pattern of connectivity (in comparison to its random start) and will have learned to perform the desired stimulus/response pairing.

2.2.2. *Connectionism and the Synthetic Approach*

How are connectionist networks related to synthetic research? First, a researcher identifies a problem of interest, and then translates this problem into some form that can be presented to a connectionist network. Second, the researcher selects a general connectionist architecture, which involves choosing the kind of processing unit, the possible pattern of connectivity, and the learning rule. Third, a network is taught the problem. This usually involves making some additional choices specific to the learning algorithm — choices about how many hidden units to use, how to present the patterns, how often to update the weights, and about the values of a number of parameters that determine how learning proceeds. If all goes according to plan, at the end of the third step the research will have constructed a network that is capable of solving a particular problem.

Connectionist networks that are built according to this general strategy are synthetic in the sense that a researcher is taking a basic set of building blocks, and is constructing a system from them in a fairly unconstrained fashion. The network does not fit existing data, but instead will create new behavior to be investigated. Specifically, while a supervised learning rule requires a researcher to dictate what a network's desired responses are, the researcher has very little control over the regularities that are used by a network to implement a mapping between input and output patterns. Indeed, one of the main surprises that can be delivered by a connectionist networks is new kinds of representations, or the discovery of new kinds of input regularities, that can be used to solve a problem, and that are completely surprising to the researcher who constructs the system.

2.2.3. *Connectionism, Representation, and Analysis*

Many of the early successes in connectionism merely involved showing that a PDP network was capable of accomplishing some task that was traditionally explained by appealing to rule-governed symbol manipulation. However, modern analyses have demonstrated conclusively that a broad variety of PDP architectures have the same computational power as the architectures that have been incorporated into symbolic accounts of cognition (Dawson, 1998). What this means is that a connectionist network can learn to perform any task that can be accomplished by a classical model. The mere fact that a network can learn a task is no longer an emergent phenomenon of any interest to researchers.

While it is neither interesting nor surprising to demonstrate that a network can learn a task of interest, it can be extremely interesting, surprising, and informative to determine what regularities the network exploits. What kinds of regularities in the input patterns has the network discovered? How does it represent these regularities? How are these regularities combined to govern the response of the network? In many instances, the answers to these questions can reveal properties of problems, and schemes for representing these properties, that were completely unexpected. This means that in order for connectionist modelers to take advantage of the emergent properties of their synthetic systems, the modelers must analyze the internal structure of the networks that they train.

Unfortunately, connectionist researchers freely admit that it is extremely difficult to determine how their networks accomplish the tasks that they have been taught (Seidenberg, 1993). Difficulties in understanding how a particular connectionist network accomplishes the task that it has been trained to perform has raised serious doubts about the ability of connectionists to provide fruitful theories about cognitive processing. Because of this, McCloskey (1991) suggested "connectionist networks should not be viewed as theories of human cognitive functions, or as simulations of theories, or even as demonstrations of specific theoretical points." Fortunately, connectionist researchers are up to this kind of challenge. Several different approaches to interpreting the algorithmic structure of PDP networks have been described in the literature (for an introduction, see Dawson, 2003, Chapter 12). The next section describes how one connectionist network was synthesized, and then analyzed, in order to contribute to our knowledge concerning one measure of higher order reasoning, Piaget's balance scale task.

3. A Case Study: The Balance Scale Task

The balance scale is considered a task of naive or intuitive physics (e.g. Wilkening and Anderson, 1982; diSessa, 1993). It has also been described as a task of 'proportional reasoning' (e.g., Kliman, 1987; Chletsos *et al.*, 1989; Normandeau *et al.*, 1989). Piaget introduced the balance scale task as a method for assessing stages of cognitive development (Inhelder and Piaget, 1958; Piaget and Inhelder, 1969). Piaget used a scale with either a sliding basket on each side of the fulcrum or 28 holes for hanging weights on each arm. Using the clinical method, Piaget allowed children of various ages to manipulate and explore the apparatus. On the basis of verbal protocols, Piaget suggested that children go through different levels of performance. By the formal operational period, children over the age of 11 or 12 were able to reason using proportions, and thereby discover the correct formula for solving balance scale problems.

3.1. PREVIOUS APPROACHES TO THE TASK

3.1.1. *The Rule Assessment Approach*

Siegler (1976) modified the balance apparatus so that there were four equidistant pegs on each side of the fulcrum. Blocks are placed under the arms to prevent the scale from tipping. The participant's task is to predict whether the scale will balance, tip to the left, or tip to the right if the blocks are removed. If one calculates *torque* (i.e., mass × distance) for each arm and compare the values, then one should be able to predict the correct response to any balance scale problem.

Siegler (1976, 1978, 1981) hypothesized that children go through a series of developmental stages in which different binary decision-tree rules are used to solve the balance scale problem. Younger children consider weight alone when deciding whether the scale will balance or not. At the next level, children focus on weight, but will consider distance information in cases where the weights are equal. Next, children realize the importance of both weight and distance, but there is some confusion when one side has the greater weight and the other side has the greater distance: performance is usually described as guessing or 'muddling through.' Lastly, the child or adult can apply the torque rule by multiplying the distance by the weight and comparing the products to determine whether the scale will balance.

Siegler (1976) described several different problem types that are defined by the combination of weight and distance from the fulcrum. *Balance* problems have equal weights at equal distances. Distance is held constant in *weight* problems, so that the side with the most weight goes down. In contrast, the weight is held constant in *distance* problems, so that the side with the farther distance goes down. Conflict problems have a different number of weights and distances on each side of the fulcrum. Three types of conflict problem were defined: *conflict-weight* (the side with more weight at a shorter distance goes down), *conflict-distance* (the side with the greater distance but with fewer weights goes down), and *conflict-balance* (despite the conflict, the scale balances).

The assessment of rule use is determined by testing the participant on a set of balance scale problems. Typically, the test set consists of four of each of the six problem types outlined above. The task is to judge which arm will tip or if the scale will balance. If a participant is using one of the four rules described above, then a characteristic pattern of performance on the different problem types should emerge (Siegler, 1976). In particular, a number of different developmental trends are predicted for the various problem types.

For instance, performance is not expected to change with age for balance or weight problems (i.e., all children should show a high level of accuracy). For distance problems, however, a dramatic improvement with age is predicted, as accuracy is expected to jump from none correct to all correct. A U-shaped trend is predicted for conflict-weight problems. Younger children should get them correct if they focus only on the weight dimension. Once children take note of distance information, performance should drop to chance level as they try to reconcile the two dimensions. When the relationship between weight and distance is understood, performance should then improve to near perfect levels. Conflict-distance and conflict-balance problems are expected to show the same pattern. Initially, performance is always incorrect; it then ‘improves’ to chance level as participants attempt to incorporate both dimensions. Finally, perfect performance follows an understanding of the multiplicative relationship between weight and distance.

3.1.2. *Symbolic Simulations of the Balance Scale Problem*

The earliest attempts to model the balance scale task used production system simulations. Klahr and Siegler (1978) used a separate set of production rules for each of the four stages of development. As there was no transition between these different rule models, the model is silent with respect to issues of stage-like development, transition mechanism, and the U-shaped trend on conflict-weight problems.

Sage and Langley (1983) created a revised model that was similar to Klahr and Siegler’s, but which included a possible mechanism to account for the transition between stages. They posited that the rules might be learned through a process of discrimination. The model was given two rules that provided initial behavior, and rules for storing information about failures and successes. New rules were acquired, and then weakened or strengthened based on the success of the predictions. Although the model never mastered conflict problems or developed a representation of torque, it learned a set of rules to make correct predictions despite the incomplete representation of the problem, just as children do. Newell (1990) developed a similar model, with similar performance, using the Soar architecture.

The most successful symbolic or rule-based model thus far has been Schmidt and Ling’s model (Schmidt and Ling, 1996) of the balance scale using Quinlan’s C4.5 machine learning algorithm (Quinlan, 1993). This algorithm generates simple decision trees that classify data that vary along a number of dimensions. For each balance problem, seven attributes were presented: (a) whether the problem was a simple balance problem (yes/no format), (b) the side with the greater weight (left, right, neither), (c) the side with the greater distance (left, right, neither), (d) left weight, (e) left distance, (f) right weight, and (g) right distance (all weight and distance values expressed as integers ranging from 1 to 5). The initial simulation displayed an orderly stage progression of all four rules when tested with Siegler’s rule-assessment technique. It also displayed a U-shaped trend on conflict-weight problems. Schmidt and Ling modified this model by using continuous values between -4 and 4 to represent attributes (b) and (c), encoding the right side minus the left side for both weight and distance, respectively. Information about weight and distance was no longer presented in an ‘all-or-none’ manner. Rather, the attribute was presented as a graded representation. Under these conditions, the model displayed all of the key regularities that were observed in psychological experiments. Schmidt and Ling concluded that their model was successful, as the selection of problem representation and learning algorithm resulted in a good match to the human data.

3.1.3. *Connectionist Models of the Balance Scale Problem*

McClelland (1989) reported the first connectionist model of balance scale phenomena. He trained a multilayer perceptron on balance scale problems that involved a scale with five pegs and a maximum of five weights on any peg. Twenty input units were used: the first 10 input units represented the left and right weight values and the remaining 10 represented left and right distance values. Inputs were segregated such that the weight and distance inputs were connected to different hidden units. A higher activation of the left or right output unit was used to indicate the side that would tip, and neutral activation of both units indicated balance.

The segregation of weight and distance information at the hidden unit level is what McClelland (1989) referred to as the *architecture assumption*. McClelland's model also included an *environment assumption*. McClelland assumed that the average child has more experience with balance scenarios in which distance is not important. Instead of training the network on the entire set of 625 possible five-peg, five-weight problems, two corpuses were developed in which there was either 5 or 10 times the equal distance problems (i.e., simple-balance and weight problems). For each training epoch, 100 patterns were randomly selected. The network was trained using a standard back-propagation learning algorithm. After each epoch, the network was tested on a 24-item test set. The network's performance was classified using Siegler's rule-assessment method (Siegler, 1976).

Without these two assumptions, McClelland's network learned the problems too quickly, often skipping the first two developmental stages, and the final developmental stage was not reliably established (Schmidt and Shultz, 1992). However, this first model did manage to demonstrate some stage-like behavior and was a fairly close fit to Siegler's predictions.

Shultz and his colleagues (Shultz and Schmidt, 1992; Shultz *et al.*, 1994, 1995) modeled balance scale phenomenon using the cascade-correlation architecture and learning algorithm (Fahlman and Lebiere, 1990). Cascade-correlation is described as a generative algorithm (Mareschal and Shultz, 1996) or one that "constructs its own network topology as it learns" (Shultz *et al.*, 1994, p. 57). The network begins without any hidden units, and then creates additional hidden units when changes in its error rates asymptote to levels that indicate that the problem has not yet been solved. To simulate the child's gradually changing environment, Shultz and Schmidt used 'expansion training,' where one new pattern was added to the training set every epoch. McClelland's architecture assumption (i.e., the segregation of weight and distance information), however, was not implemented in this model.

Sixteen "computer subjects" were used in the Shultz and Schmidt experiment (i.e., the simulation was run 16 separate times with different initial weight states). After each training epoch, the network was tested on 24 randomly selected test patterns. Rule diagnosis was similar to that used by Siegler (1976) for human subjects. With respect to matching the developmental sequence, the majority of the computer subjects progressed through all four stages in the appropriate order. Most of the remaining simulations progressed through stages in the correct order, but did not achieve the final stage of competence with the balance scale task. These same patterns are found with human subjects (e.g., Siegler, 1981; Chletsos *et al.*, 1989).

Shultz *et al.* (1995) reported a second cascade-correlation model of the balance scale. Rather than biasing the training environment toward weight information, the internal state of the network was prestructured so that preferential treatment would be given to the weight information. The entire set of 625 possible problems was presented during training. The network displayed the regularities found in human performance, including the use of all four of Siegler’s rules in the correct order, some stage skipping and regressions and U-shaped development on conflict-weight problems.

3.2. A SYNTHETIC APPROACH TO THE BALANCE SCALE PROBLEM

3.2.1. *Problems With Previous Analytic Approaches*

To date, models of the balance scale problem have been analytic, because they have been created in the service of fitting the psychological data. For each of the models discussed above, one of the primary goals of the researcher was to generate data that conformed to the pattern of experimental results of the sort reported by Siegler (1976). For example, Schmidt and Ling (1996) commented that, “Regardless of the learning algorithm that one adopts (connectionist or symbolic), the choice of attributes to use is crucial *if the model’s output is to match the human data* (p. 211, emphasis added).

One major problem with this approach is that, ultimately, a model’s match to the human data is determined by its fit to Siegler’s rule assessment approach (Siegler, 1976). This is an issue because rule assessment has been subject to a number of criticisms (e.g., Wilkening and Anderson, 1982; Ferretti *et al.*, 1985; Flavell, 1985; Ferretti and Butterfield, 1986; Kliman, 1987; Laviree *et al.*, 1987; Chletsos *et al.*, 1989; Normandeau *et al.*, 1989; van Maanen *et al.*, 1989; Jansen and van der Mass, 1997) The criticisms that have been leveled against rule assessment include (a) arbitrary criteria for scoring, (b) assessment varying with the torque difference of the items used, (c) assessment varying with the priority given to the various rules, (d) assessment varying with task demands, (e) scoring criteria that are not diagnostic with respect to other postulated rules, and (f) lack of clarity regarding the “muddle through” stage. These critiques, cumulatively, make the interpretation of data from psychological studies either doubtful or ambiguous. They in turn pose questions for studies that attempt to compare model outputs to rule assessment data.

One approach to dealing with this problem would be to adopt the synthetic perspective advocated by Dawson (2003). Rather than designing a model to fit human data, it is plausible to synthesize a model that is capable of solving the balance scale task. The task can then be explored by interpreting the internal structure of this model. This approach is described in the sections that follow.

3.2.2. *Synthesis of a Network for the Balance Scale Task*

3.2.2.1. Network architecture. We trained a multilayer perceptron to solve the balance scale problem. The network had 20 input units and 4 hidden units, and was fully connected in that every input unit was connected to every hidden unit (i.e., the input was not segregated as in McClelland, 1989). Two output units were used. Activation of either the left or right output unit represented the corresponding side of the balance scale that would tip. Problems in which the scale balanced were represented by zero values on both output units. All hidden units and output units were *value units* (Dawson and Schopflocher, 1992), which use a Gaussian activation function. Value units were selected because value units have been shown to be particularly interpretable (Dawson, 2003).

3.2.2.2. Training set. The network was presented with all 625 possible five-peg, five-weight problems. balance scale problems were represented on 20 input units. The first five units represented left weight and were thermometer coded. In this coding, a unit was turned on for every weight that was placed on a peg. So, if two weights were on a peg, the first two input units were activated; if four weights were on a peg, then the first four input units were activated. The second set of five units represented left distance and were unary coded. With this coding, only one unit is turned on to represent which peg the weights were placed on. For example, if the third peg were to be used, then only the third input unit in this group would be activated. The same encoding was repeated for the remaining 10 input units to represent right weight and right distance. The desired response of the network was determined by applying the torque rule for each of the 625 input patterns.

3.2.2.3. Training the network. The network was trained using Dawson and Schopflocher's version of the generalized delta rule designed for use with value units (Dawson and Schopflocher, 1992). The learning rate was 0.005, and no momentum term was used. Connection weights were randomly set in the range of ±0.1 and the biases of all value units (i.e., the mean for the Gaussian for each unit) were initialized to zero. In pilot tests, these parameter settings resulted in reliable convergence. The network was trained until a 'hit' was obtained on every output unit for every input pattern. The criterion for a 'hit' was set at 0.01 (i.e., an output unit's activity had to be greater than or equal to 0.90 when the desired output was 1, and less than or equal to 0.10 when the desired output was 0). Pattern order presentation was randomized every epoch. Network connections and biases were updated after each pattern presentation. The network converged to a solution to the balance scale problem after 4120 epochs of training.

3.2.3. *Interpretation of Hidden Unit Activations*

At the end of training, the network was capable of making the correct balance scale judgment for each of the 625 problems. That is, its input/output behavior was as if the network was using its internal structure to compute torque. How exactly were the hidden units being used to solve this problem?

One possibility was that each hidden unit was computing a different aspect of the torque equation. For instance, it was plausible that one hidden unit was computing the left weight, a second was computing the right weight, and that the remaining hidden units were each computing the left and right distance. However, an examination of the connection weights feeding into each hidden unit, and an examination of each hidden unit's response to the input patterns, indicated that the problem was *not* being solved in this manner. There was no evidence that individual hidden units were tuned to weight or to distance alone. Instead, each hidden unit appeared to be generating a response that was sensitive to a combination of both weight and distance.

How were the hidden units combining weight and distance? The hidden units could not be explicitly computing torque, because this would require signals from different input units (i.e., units representing weight and units representing distance) to be multiplied together. Multiplication is not a primitive operation in the net input function for a value unit. The only way that the hidden units could combine weight and distance would be to add them together in some manner. This suggests that the hidden units might be computing some additive rule, such as the equation [RW + RD] − [LW + LD], which is one of the alternative rules for the balance scale problem suggested by Wilkening and Anderson (1982).

In order to test the possibility that the hidden units might be approximating the torque rule with an additive equation, the net input for each hidden unit was computed for each of the 625 input patterns. These net inputs were then correlated with the torque equation ([LW × LD] − [RW × RD]) and with the additive rule ([RW + RD] − [LW + LD]). For the torque equation, the correlations obtained for hidden units 1 through 4 were 0.92, 0.92, −0.87, and −0.92. For the additive rule, the four correlations were even higher: 0.97, 0.97, −0.92, and −0.97. What this suggests is that the net inputs being computed were almost perfectly related to the additive rule, which in turn provides a very strong approximation to the torque rule.

Why, then, does the network require four hidden units to compute the same function? The answer is that while each hidden unit appears to be computing the additive rule, it has only a limited sensitivity to the result of the calculation when it is converted into internal activity by the Gaussian equation. Four hidden units are required because each hidden unit is sensitive to different range of additive rule results.

This can be seen in Figure 1, which illustrates the activity of each hidden unit as a function of the additive rule result for each problem in the training set. When the value returned by the additive rule is extreme, a single hidden unit responds (i.e., Hidden units 1 and 3 for left and right problems, respectively). When the value returned is nearer to

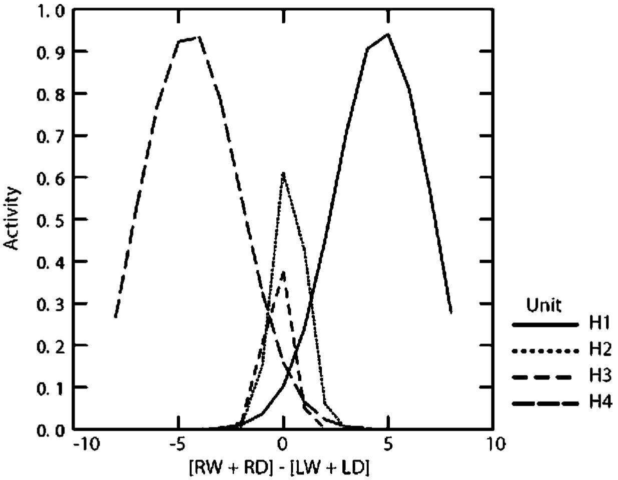

Figure 1. Activation of the four hidden units as a function of the additive equation computed for each of the patterns in the training set.

zero, the overlapping sensitivities of all four hidden units are required to make the correct prediction. This situation is analogous to the use of overlapping receptive fields in the visual system to make fine spatial discriminations, and is a representational scheme called *coarse coding* (Hinton *et al.*, 1986).

3.2.4. *Exploration of the Network's Pattern Space*

In interpreting the internal structure of a network, examining the role of individual hidden units is just one step. A second step involves examining how the activities of different hidden units are combined to determine the network's output. In value unit networks, when coarse coding is discovered, this second step is typically accomplished by performing a cluster analysis of hidden unit activities (for details, see Dawson, 2003, Chapter 12). In general, *k*-means cluster analysis is applied to the set of activities of all of the hidden units for all of the input patterns in order to identify pattern regularities that are distributed across hidden units. In other words, the result of this analysis is a clustering of different input patterns, based on the similarities of activation patterns that they produce in the entire set of a network's hidden units. The number of clusters that patterns are assigned to is determined by a heuristic stopping rule: we choose the smallest number of clusters such that each pattern that falls into the same cluster maps onto the same output response in the network.

Using this stopping heuristic, the input patterns were assigned to seven different clusters. Three clusters (1, 6, and 7) contained *left* problems, three (2, 4, and 5) contained *right* problems, and one cluster (3) contained all *balance* problems. The representation of these clusters in a pattern space for the network is provided in Figure 2. The pattern space for the training set cannot be shown properly in four dimensions, and so these figures illustrate the pattern space in two dimensions by plotting patterns as a function of left torque (i.e., LW × LD) and right torque (i.e., RW × RD).

It is apparent from Figure 2 that the problems associated with each cluster fall into a different, nonoverlapping region of the pattern space. Most of the regions are in the shape of a relatively narrow triangle, and all are extended through the space. This kind of pattern space is ideally suited for classification by value units. This is because value units carve two parallel cuts through a pattern space, and every pattern that falls between the cuts will turn the output unit on. By arranging the orientation of the cuts in such a way that only certain clusters of problems fall between them, the output units could easily use this pattern space to generate correct responses to the balance scale problem. In short, the different sensitivities of the four hidden units to the additive rule create a pattern space in which seven different types of problems are arranged in a regular fashion that permits the output units to correctly generate a problem response.

Further to this, there is a pattern of hidden unit activity that characterizes each cluster. For example, hidden unit 1 (H1) is associated with left patterns. Cluster 1 includes the patterns with large torque differences (TD) and strong activation of H1. Cluster 7 includes patterns with a medium level TD and a medium-level activation of H1. The same pattern holds for right patterns, with H3 having strong and moderate levels of activation for patterns in Clusters 2 and 5, respectively. A slightly different pattern emerged for patterns with low

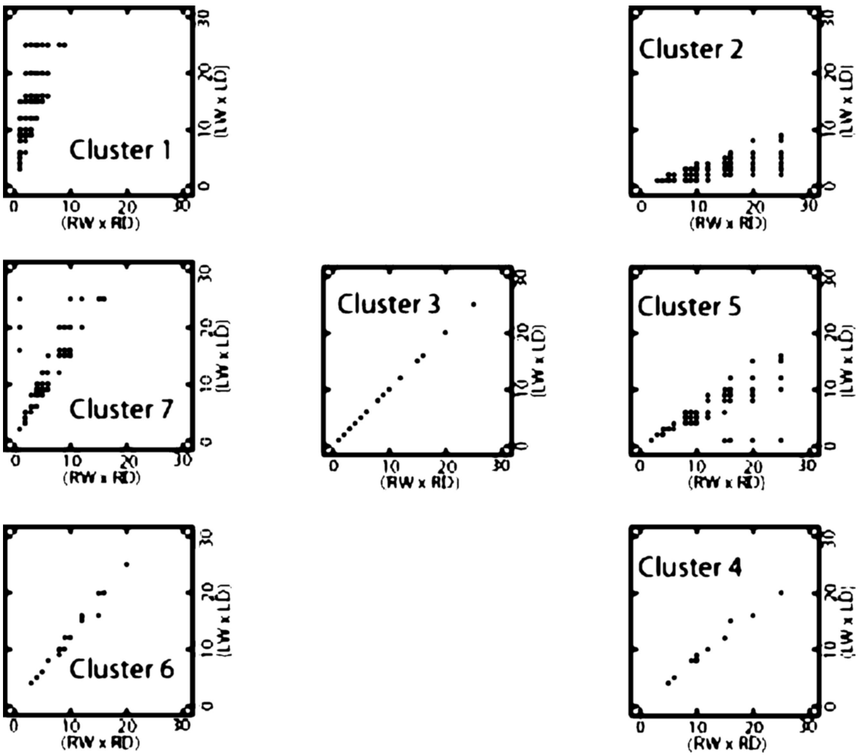

Figure 2. Location of input patterns for each of seven clusters in a pattern space defined by torque difference. For each graph, the x-axis is right torque, and the y-axis is left torque. Each graph represents the locations of patterns in the space where every pattern in the graph belongs to the same cluster.

TD. For left and right patterns with low TD, there were rather high activations for H2 and H4 respectively, along with low levels of activation distributed across the other two hidden units. Similarly, balance problems (i.e., TD = 0) were associated with distributed hidden unit activities. In general then, the network was sensitive to differences in torque. This kind of sensitivity was found by Ferretti and Butterfield (1986) in human subjects, and was found independently by Shultz and Schmidt (1992) using cascade-correlation networks.

The cluster analysis reported above provides a classification of different types of balance scale problems. We have already seen balance scale problems classified in a different theory, the rule assessment approach of Siegler (1976). What is the relationship between the two classification schemes?

If one were to use Siegler’s taxonomy (Siegler, 1976) to generate problem space plots similar to those in Figure 2, then one would find that this space would not lend itself very well to solving the balance scale problem by carving the pattern space into different regions. This would only work for two of his problem types, balance and conflict balance. This is because the balance problems would fall on the diagonal of the pattern space, similar to Cluster 3 in Figure 2. However, the other four problem types are smeared in overlapping regions across much of the pattern space. As a result, their position in the pattern space could not be used to make a correct response, because problems (defined by Siegler's taxonomy) that lead to very different responses occupy similar locations in the space.

To make this point mathematically, for each of Siegler's problem types, one can correlate its left torque value with its right torque value to determine how constrained its region in the pattern space would be. For the two types of balance problems in his taxonomy, the correlations are 1.00. However, for the other four types of problems in his taxonomy, the correlations range from only 0.29 to 0.31. In contrast, if one were to compute the same correlations for the clusters illustrated in Figure 2, they are much higher, ranging from 0.64 to 0.99 for clusters involving problems other than those that balance. Of course the correlation for balance problems is 1.00. In short, the problem types identified by analyzing the network are quite different from those proposed by Siegler, and occupy much more constrained regions in the pattern space than would patterns organized by his classification scheme.

3.3. IMPLICATIONS

The sections above have described a synthetic approach to the balance scale task. In contrast to previous approaches, which have primarily attempted to develop models that fit the data predicted by Siegler's rule assessment approach (Siegler, 1976), we simply constructed a network that was capable of solving the balance scale task. Then we interpreted its internal structure, with the intent of discovering surprising emergent properties that could inform us about the nature of this task, and hopefully provide insights that could be incorporated in the study of how children might approach this task. What new insights has this synthetic approach revealed?

3.3.1. *A New Rule for the Balance Scale Task*

In most previous studies of the balance scale task, it has been assumed that perfect performance is defined by a multiplicative torque rule. There has been little work that has explored alternative approaches to defining solutions to the problem (for an exception, see Wilkening and Anderson, 1982). In the analysis of the network above, we were forced to consider alternative equations for solving the problem, because (a) it was apparent that each hidden unit was computing a function that combined weight and distance, and (b) constraints on the processes that are present in the hidden units prevented the traditional torque equation from being this function. Our analysis revealed that hidden units were computing an additive rule ($[RW + RD] - [LW + LD]$), that this rule was capable of generating the correct solution to each balance scale problem in the training set, and that this rule was an excellent approximation to the more traditional torque rule. It is important to stress that the only reason that this rule was discovered was because of our need to interpret the internal structure of the network after it was trained. We did not (analytically) attempt to build a network that computed this function; rather, we (synthetically) built a network that solved the problem, and were informed about the existence of this rule by exploring the structure of the system that we had constructed.

3.3.2. *A New Taxonomy of Balance Scale Problems*

The rule assessment approach of Siegler (1976) is based on a decision-tree theory in which particular rules are applied in a particular order. One consequence of this theory has been the proposal of a particular taxonomy of balance scale problems. Siegler proposed that there were six different classes of problems (for more details, see Section 3.1.1), and his theory predicted different courses of development for the different types of problems.

Siegler's taxonomy (Siegler, 1976) has been central to the experimental study of the balance scale task. It is not possible to test children on all possible problems. Instead, the assessment of rule use is determined by testing the participant on a set of balance scale problems. Typically, the test set consists of a small number of each of the six problem types that were outlined earlier. In order to be consistent with data from human experiments, the evaluation of many of the simulation studies that were reviewed above follows a similar practice. For example, Shultz and Schmidt (1992) used this exact procedure to apply the rule assessment theory to their cascade correlation network.

The application of Siegler's taxonomy and rule assessment method to computer models of the balance scale task is understandable (Siegler, 1976), but it neglects to address one extremely interesting issue: given all the criticisms of the rule assessment, are there other plausible taxonomies of problems that could be used to develop an alternative theory of the balance scale task? Our cluster analysis of hidden unit activations can be interpreted as providing an alternative taxonomy of balance scale problems. The network that was described above was *not* responding to the different types of problems as defined by Siegler's taxonomy. Instead, the network was arranging the problems in a pattern space that defines one general problem characteristic, torque difference. Furthermore, there were seven distinct regions within this pattern space, and each can be considered as defining a new category type for the balance scale problems that fall into this region. This new approach to classifying balance scale problems would never have been revealed had we adopted an analytic approach, and attempted to fit the behavior of our network to Siegler's preexisting theory that the task is solved via a series of binary decision-tree rules.

3.3.3. *New Predictions for Experimental Study*

In previous work on the balance scale task (e.g., Siegler, 1976; McClelland, 1989; Shultz and Schmidt, 1992) the problem types defined by Siegler are essential for determining the rule used (by humans or computer models) to solve balance problems. The results of the network analysis yield a number of novel predictions to be contrasted with Siegler's theory. For example, the network responded differentially to problems based on a characteristic of the problem: *torque difference* (the absolute difference between the torque on the left and right side). One prediction for human performance then, is that reaction times (time to make a prediction) should vary as a function of the torque difference of a problem. In contrast, the decision-tree theory suggests that a number of properties must be considered prior to making a prediction and therefore RT differences are predicted for the different *types* of problems, but is silent with respect to the torque difference of a problem (e.g., all weight problems require fewer decisions than all conflict-weight problems and should be solved faster, regardless of torque difference). This, and other predictions for performance, were outlined and tested with a large set of balance scale problems (Zimmerman, 1999), with the results for adults matching the predictions derived from the network analysis. Pilot studies with younger participants (age 8) show the same performance trends as adults. That is, it was found that RT varied as a function of TD (which is closely correlated with cluster membership) but not as a function of Siegler's problem types.

4. General Implications

The preceding case study demonstrates that one can use connectionism to conduct a synthetic psychology that is representational, and use the interpretation of the internal structure of a network to contribute to our understanding of balance scale phenomena. This synthetic approach suggested an alternative way that the balance scale task could be solved (an additive heuristic rather than binary decision rules), revealed additional evidence for the importance of a particular characteristic of balance problems (i.e., torque difference) for making predictions, and provided empirically testable predictions for human performance. At this early stage in this research program, we are not yet ready to decide what problem domains are well suited to the kind of approach that was illustrated above. General claims about the limitations of this synthetic methodology await future research. However, we have also used this approach to contribute to a wide variety of other psychological domains, including solving logic problems (Dawson *et al.*, 1997), deductive and inductive reasoning (Leighton and Dawson, 2001), spatial reasoning (Dawson *et al.*, 2000), and the relation between symbolic and subsymbolic theories of mind (Dawson and Piercey, 2001). To the extent that researchers are eager to explore alternative representational approaches to a variety of cognitive problems, we would suggest that this synthetic use of connectionist modeling could make important contributions to the field of cognitive informatics.

Acknowledgments

This research was supported by NSERC Discovery Grant 2038-02 and by SSHRC Research Grant 410-2002-0351, both awarded to M.R.W.D.

References

Anderson, J. R. and Bower, G. H., 1973: *Human Associative Memory*, Erlbaum, Hillsdale, NJ.

Ashby, W. R. (1960). *Design for a Brain*, 2nd edn., Wiley, New York.

Bever, T. G., Fodor, J. A. and Garrett, M., 1968: A formal limitation of associationism, in T. R. Dixon and D. L. Horton (eds.), *Verbal Behavior and General Behavior Theory*, Prentice-Hall, Englewood Cliffs, NJ, pp. 582–585.

Braitenberg, V., 1984: *Vehicles: Explorations in Synthetic Psychology*, MIT Press, Cambridge, MA.

Brooks, R. A., 1999: *Cambrian Intelligence: The Early History of the New AI*, MIT Press, Cambridge, MA.

Chletsos, P. N., de Lisi, R., Turner, G. and McGillicuddy-de Lisi, A. V., 1989: Cognitive assessment of proportional reasoning strategies, *J. Res. Dev. Edu.* **22**, 18–27.

Chomsky, N., 1959: A review of B.F. Skinner's Verbal Behavior, *Language* **35**, 26–58.

Chomsky, N., 1965: *Aspects of the Theory of Syntax*, MIT Press, Cambridge, MA.

Chomsky, N., 1995: *The Minimalist Program*, MIT Press, Cambridge, MA.

Chomsky, N. and Halle, M., 1991: *The Sound Pattern of English*, MIT Press, Cambridge, MA.

Dawson, M. R. W., 1998: *Understanding Cognitive Science*, Blackwell, Oxford.

Dawson, M. R. W., 2003: *Minds and Machines: Connectionism and Psychological Modeling*, Blackwell, Oxford.

Dawson, M. R. W., Boechler, P. M. and Valsangkar-Smyth, M, 2000: Representing space in a PDP network: Coarse allocentric coding can mediate metric and nonmetric spatial judgments, *Spat. Cogn. Comput.* **2**, 181–218.

Dawson, M. R. W., Medler, D. A. and Berkeley, I. S. N., 1997: PDP networks can provide models that are not mere implementations of classical theories, *Philos. Psychol.* **10**, 25–40.

Dawson, M. R. W. and Piercey, C. D., 2001: On the subsymbolic nature of a PDP architecture that uses a nonmonotonic activation function, *Minds Machines* **11**, 197–218.

Dawson, M. R. W. and Schopflocher, D. P., 1992: Modifying the generalized delta rule to train networks of nonmonotonic processors for pattern classification, *Connect. Sci.* **4**, 19–31.

diSessa, A. A., 1993: Toward an epistemology of physics, *Cogn. Inst.* **10**, 105–225.

Fahlman, S. E. and Lebiere, C., 1990: *The Cascade-Correlation Learning Algorithm* (CMU-CS-90-100), School of Computer Science, Carnegie Mellon University, Pittsburgh.

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

Ferretti, R. P., Butterfield, E. C., Cahn, A. and Kerkman, D., 1985: The classification of children's knowledge: Development on the balance scale and inclined-plane tasks. *J. Exp. Child Psychol.* **39**, 131–160.

Flavell, J. H., 1985: *Cognitive Development*, Prentice-Hall, Englewood Cliffs, NJ.

Fodor, J. A., 1975: *The Language of Thought*, Harvard University Press, Cambridge, MA.

Grey Walter, W., 1950: An imitation of life, *Sci. Am.* **182** (5), 42–45.

Grey Walter, W., 1951: A machine that learns, *Sci. Am.* **184**(8), 60–63.

Grey Walter, W., 1963: *The Living Brain*, W.W. Norton & Co., New York.

Hinton, G. E., McClelland, J. and Rumelhart, D., 1986: Distributed representations, in D. Rumelhart and J. McClelland (eds.), *Parallel Distributed Processing, Vol. 1*, MIT Press, Cambridge, MA. pp. 77–109.

Inhelder, B. and Piaget, J., 1958: *The Growth of Logical Thinking From Childhood to Adolescence*, Basic Books, New York.

Jackendoff, R., 1992: *Languages of the Mind*, MIT Press, Cambridge, MA.

Jansen, B. R. J. and van der Mass, H. L. J., 1997: Statistical test of the rule assessment methodology by latent class analysis, *Dev. Rev.*, **17**, 321–357.

Klahr, D. and Siegler, R. S., 1978: The representation of children's knowledge, in H. W. Reese and L. P. Lipsitt (eds.), *Advances in Child Development and Behavior, Vol. 12*, Academic Press, New York, pp. 61–116.

Kliman, M., 1987: Children's learning about the balance scale. *Inst. Sci.* **15**, 307–340.

Laviree, S., Normandeau, S., Roulin, J. L. and Longeot, F., 1987: Siegler's balance scale: A critical analysis of the rule-assessment approach, *L'Anee Psychol.* **87**, 509–534.

Leighton, J. P. and Dawson, M. R. W., 2001: A parallel distributed processing model of Wason's selection task, *Cogn. Syst. Res.* **2**, 207–231.

Mareschal, D. and Shultz, T. R., 1996: Generative connectionist networks and constructivist cognitive development, *Cogn. Dev.* **11**, 571–603.

Marr, D., 1982: *Vision*, W.H. Freeman, San Francisco.

McClelland, J., 1989: Parallel distributed processing: Implications for cognition and development, in R. G. M. Morris (ed.), *Parallel Distributed Processing: Implications for Psychology and Neurobiology*, Oxford University Press, Oxford, pp. 8–45.

McCloskey, M., 1991: Networks and theories: The place of connectionism in cognitive science, *Psychol. Sci.* **2**, 387–395.

Minsky, M. and Papert, S., 1988: *Perceptrons, 3rd edn.*, MIT Press, Cambridge, MA.

Moravec, H., 1999: *Robot*, Oxford University Press, New York.

Newell, A., 1990: *Unified Theories of Cognition*, Harvard University Press, Cambridge, MA.

Normandeau, S., Laviree, S., Roulin, J. L. and Longeot, F., 1989: The balance-scale dilemma: Either the subject or the experimenter muddles through, *J. Genet. Psychol.* **150**, 237–249.

Paivio, A., 1969: Mental imagery in associative learning and memory, *Psychol. Rev.* **76**, 241–263.

Paivio, A., 1971: *Imagery and Verbal Processes*, Holt, Rinehart and Winston, New York.

Pfeifer, R. and Scheier, C., 1999: *Understanding Intelligence*, MIT Press, Cambridge, MA.

Piaget, P. and Inhelder, B., 1969: *The Psychology of the Child*, Routledge and Kegan Paul, London.

Pylyshyn, Z. W., 1984: *Computation and Cognition*, MIT Press, Cambridge, MA.

Quinlan, J. R., 1993: *C4.5: Programs for Machine Learning*, Morgan Kaufmann, San Mateo, CA.

Rumelhart, D. E., Hinton, G. E. and Williams, R. J., 1986: Learning representations by back-propagating errors. *Nature* **323**, 533–536.

Sage, S. and Langley, P., 1983: Modeling cognitive development on the balance scale task, in *Proc Eighth International Joint Conference on Artificial Intelligence, Vol. 1*, 94–96.

Schmidt, W. C. and Ling, C. X., 1996: A decision-tree model of balance scale development, *Machine Learn.* **24**, 203–230.

Schmidt, W. C. and Shultz, T. R., 1992: An investigation of balance scale success in Paper, in *Presented at the Fourteenth Annual Conference of the Cognitive Science Society*.

Seidenberg, M., 1993: Connectionist models and cognitive theory. *Psychol. Sci.* **4**, 228–235.

Shultz, T. R., Mareschal, D. and Schmidt, W. C., 1994: Modeling cognitive development on balance scale phenomena, *Machine Learn.* **16**, 57–86.

Shultz, T. R. and Schmidt, W. C., 1992: A cascade-correlation model of balance scale phenomena, in *Paper, Presented at the Thirteenth Annual Conference of the Cognitive Science Society*.

Shultz, T. R., Schmidt, W. C., Buckingham, D. and Mareschal, D., 1995: Modeling cognitive development with a generative connectionist algorithm, in T. J. Simon and G. S. Halford (eds.), *Developing Cognitive Competence: New Approaches to Process Modeling*, Erlbaum Hillsdale, NJ, pp. 205–261.

Siegler, R. S., 1976: Three aspects of cognitive development, *Cogn. Psychol.* **8**, 481–520.

Siegler, R. S., 1978: The origins of scientific reasoning, in R. S. Siegler (ed.), *Children's Thinking: What Develops?* Erlbaum, Hillsdale, NJ.

Siegler, R. S., 1981: Developmental sequences within and between concepts. *Monogr. Soc. Res. Child Dev.* **46**(Whole No. 189).

Skinner, B. F., 1957: *Verbal Behavior*, Appleton-Century-Crofts, New York.

Smolensky, P., 1988 : On the proper treatment of connectionism, *Behav. Brain Sci.* **11**, 1–74.

van Maanen, L., Been, P. and Sijtsma, K., 1989: The linear logistic test model and heterogeneity of cognitive strategies, in E. E. Roskam (ed.), *Mathematical Psychology in Progress*, Springer-Verlag, Berlin, (pp. 267–287).

Watson, J. B., 1913: Psychology as the behaviorist views it. *Psychol. Rev.* **20**, 158–177.

Wilkening, F. and Anderson, N. H., 1982: Comparison of two rule-assessment methodologies for studying cognitive development and knowledge structures, *Psychol. Bull.* **92**, 215–237.

Zimmerman, C. L., 1999: *A network interpretation approach to the balance scale task*, Unpublished PhD, Thesis, University of Alberta, Edmonton, Canada.