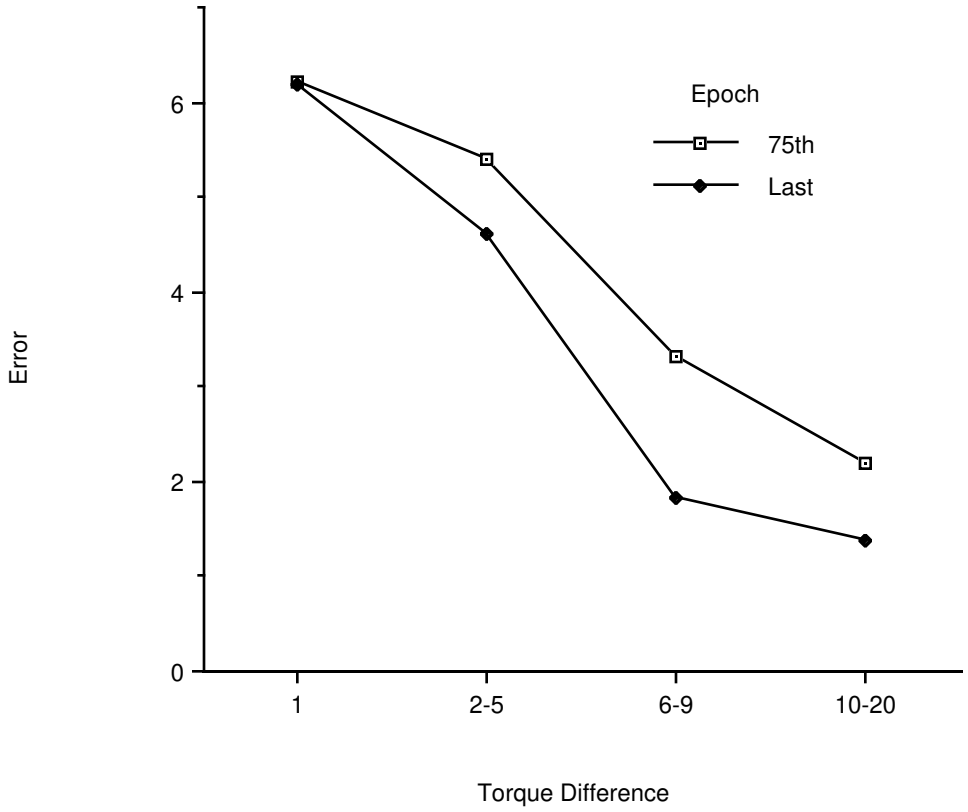


Figure 10. Mean errors on balance scale test problems at four torque difference levels.

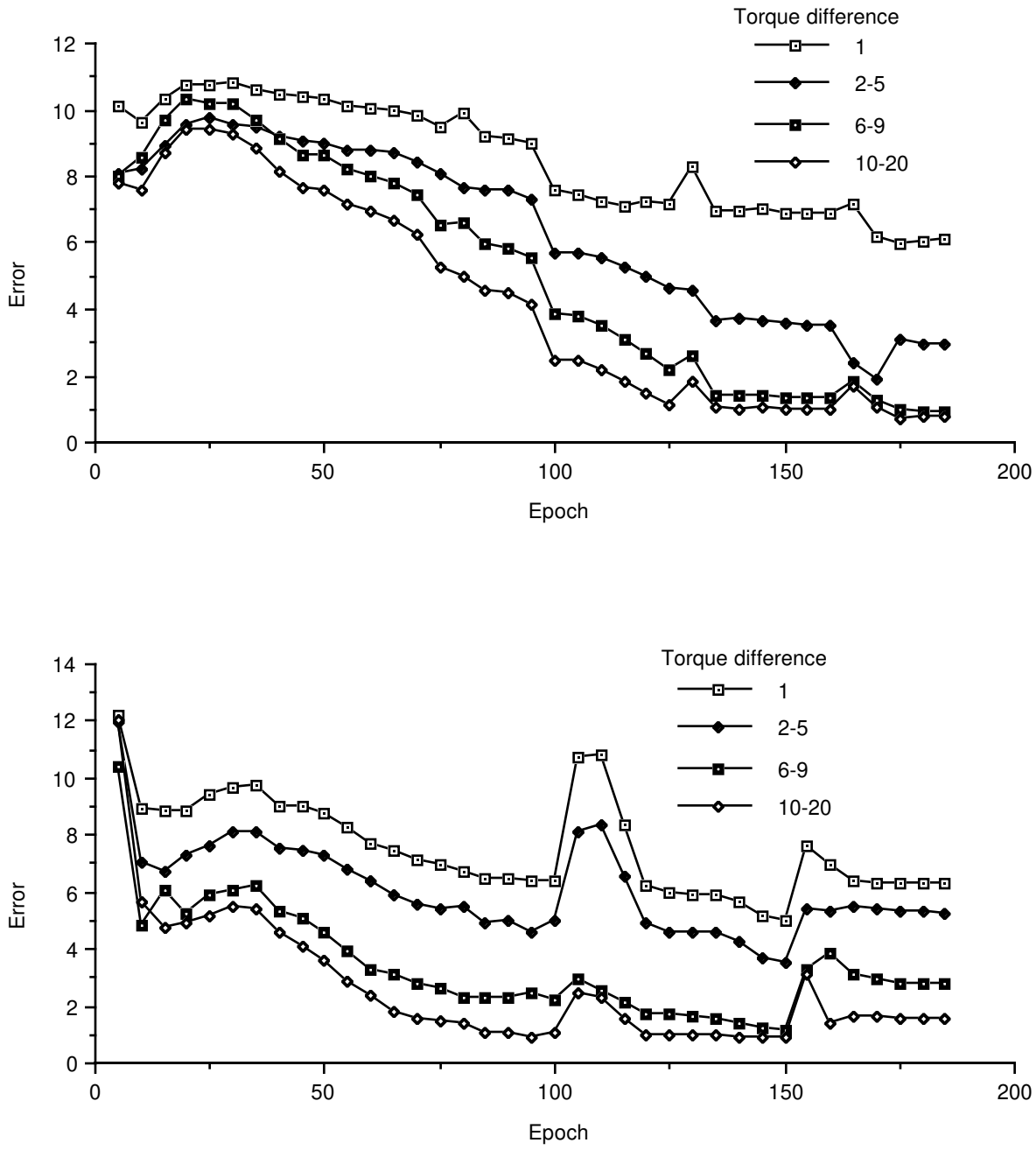


Figure 9. Errors on test problems at four levels of torque difference, plotted every fifth epoch, for two distinct networks.

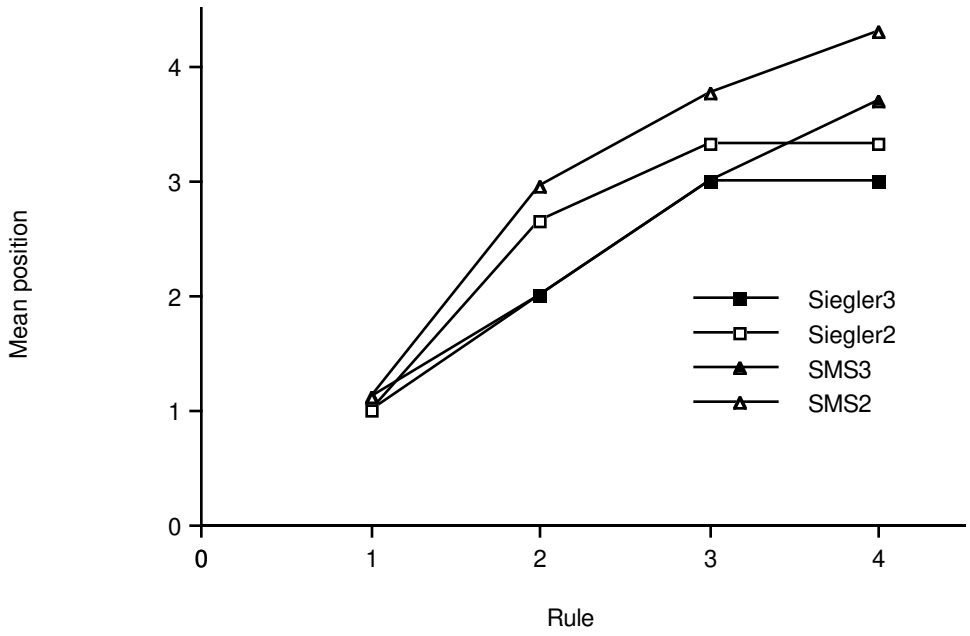


Figure 8. Mean position of each rule in stages as diagnosed by each of four scoring methods.

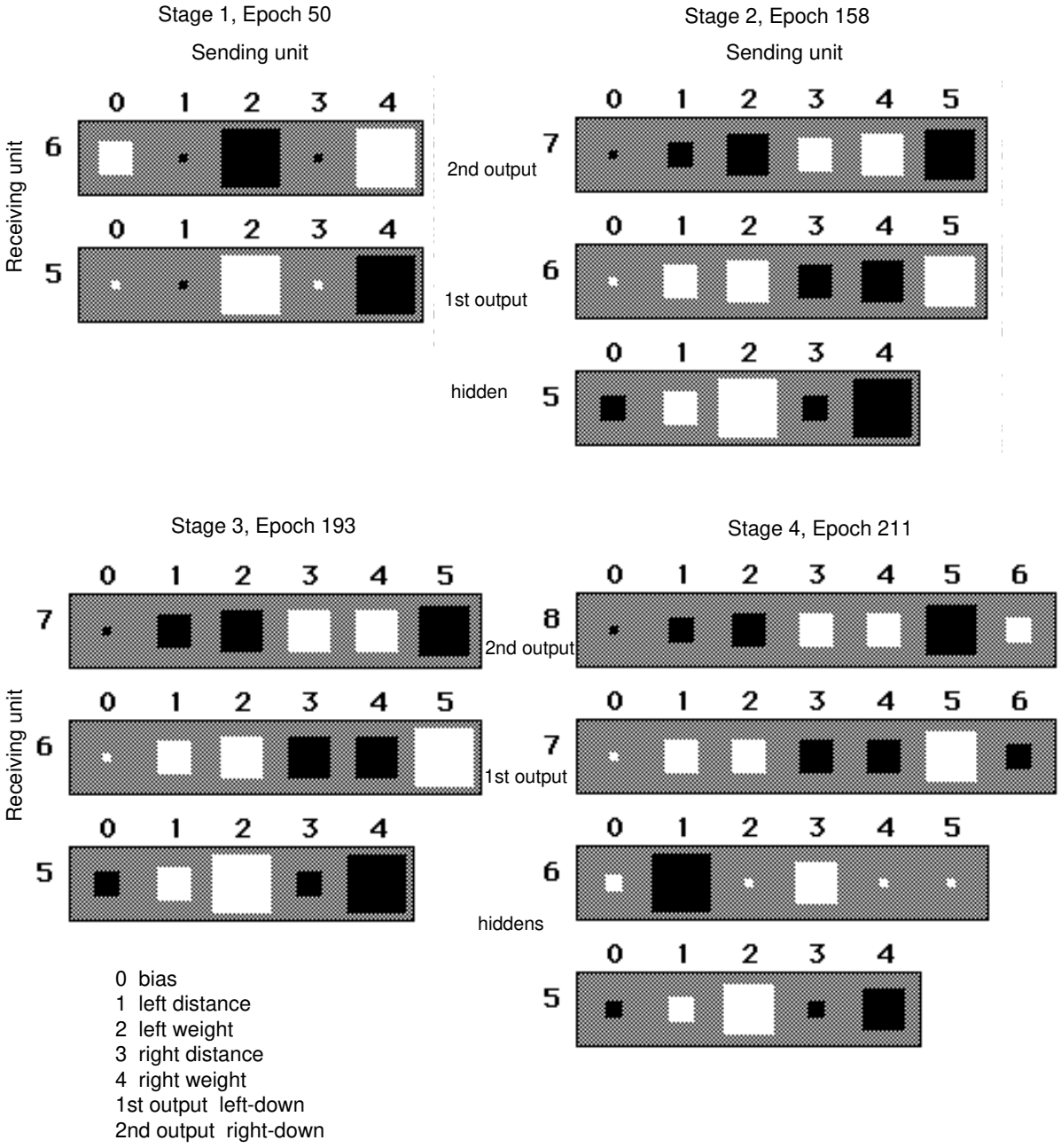


Figure 7. Hinton diagrams of incoming weights in the middle of each of four stages for network 16.

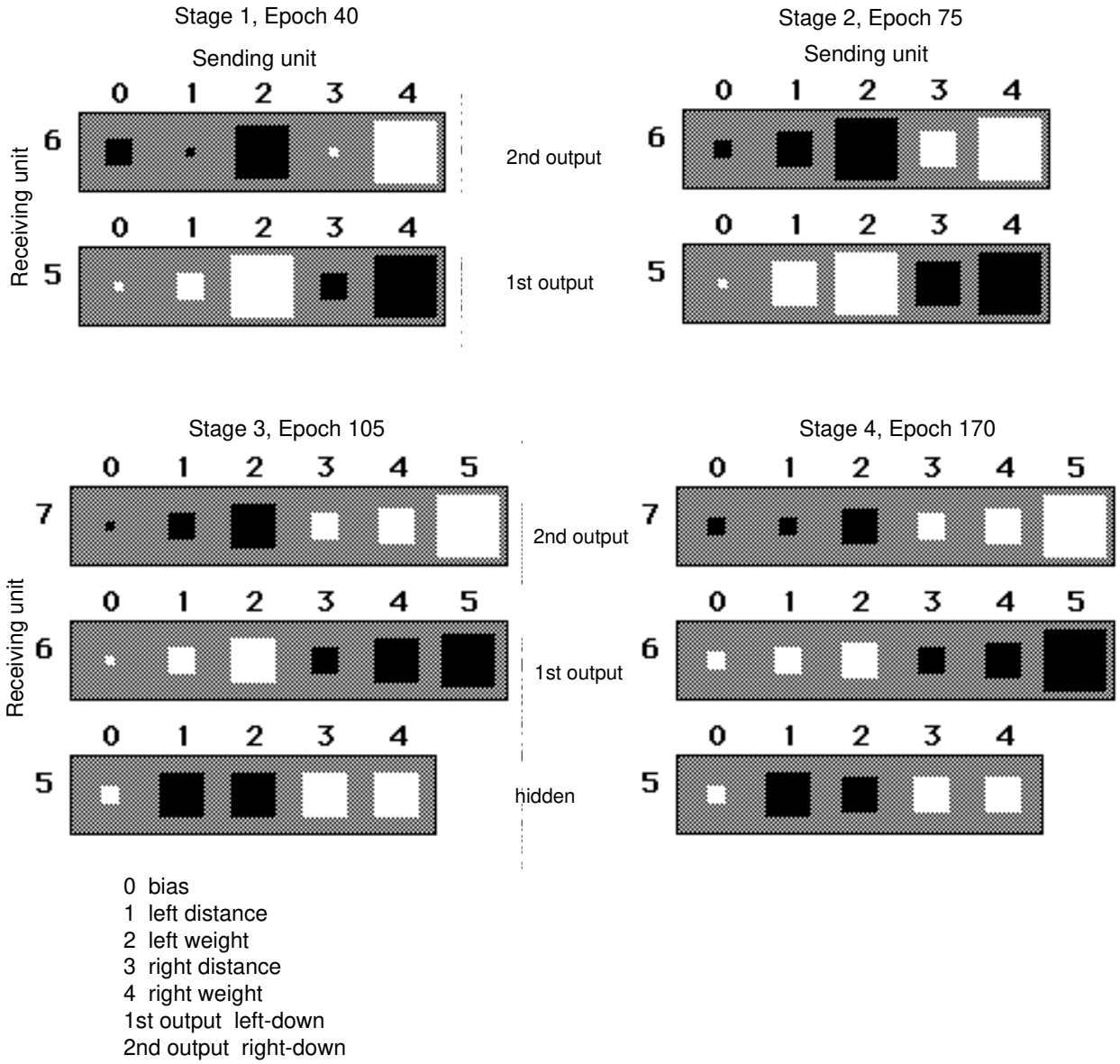


Figure 6. Hinton diagrams of incoming weights in the middle of each of four stages for network 1.

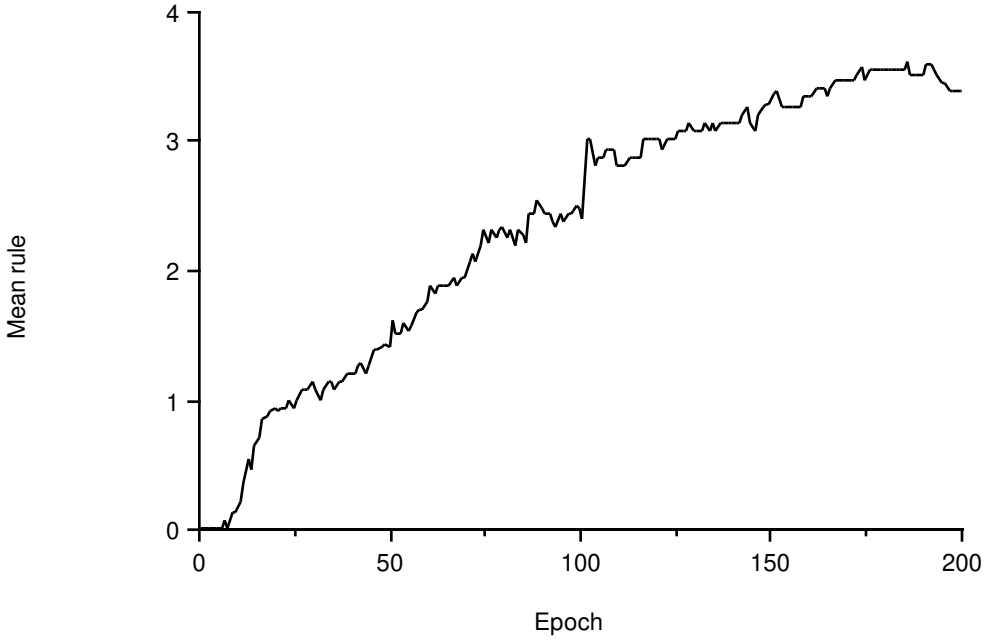


Figure 5. Mean rule diagnosis for all 16 nets on the balance scale task.

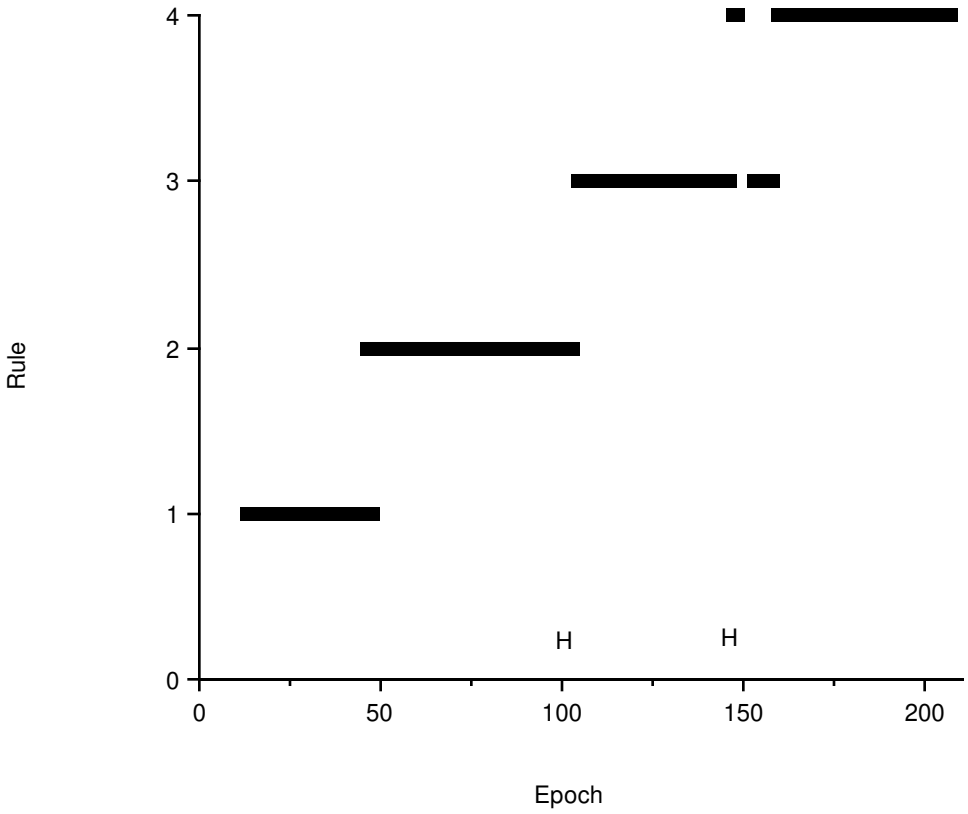
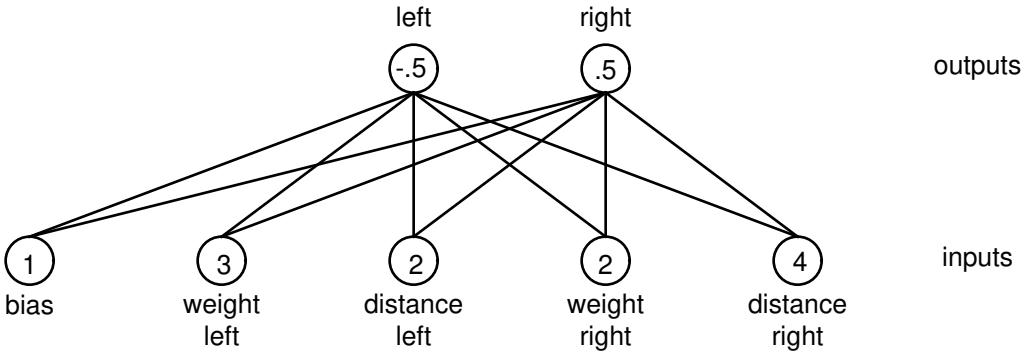
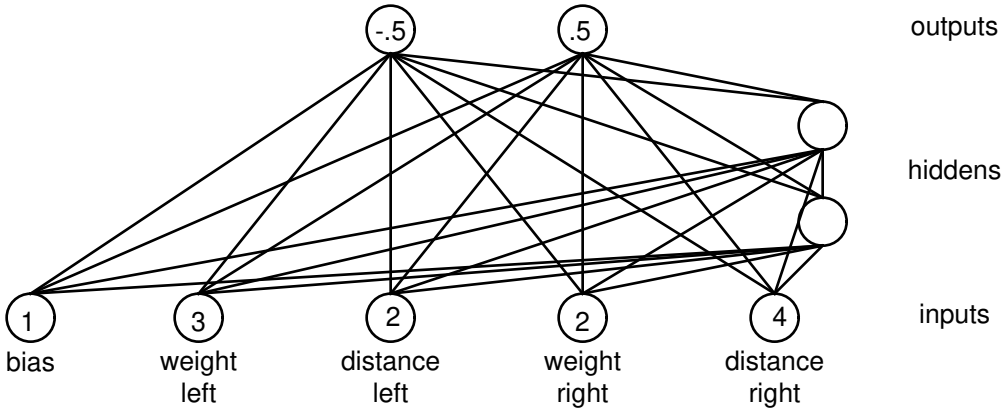


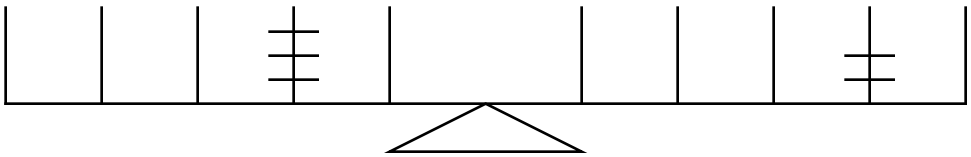
Figure 4. Rule diagnosis for one net on the balance scale task. H indicates the epochs at which hidden units were installed.



a.



b.



c.

Figure 3. Initial network (a) and network after addition of two hidden units (b). Activation values represent the input and output coding for balance scale problem (c).

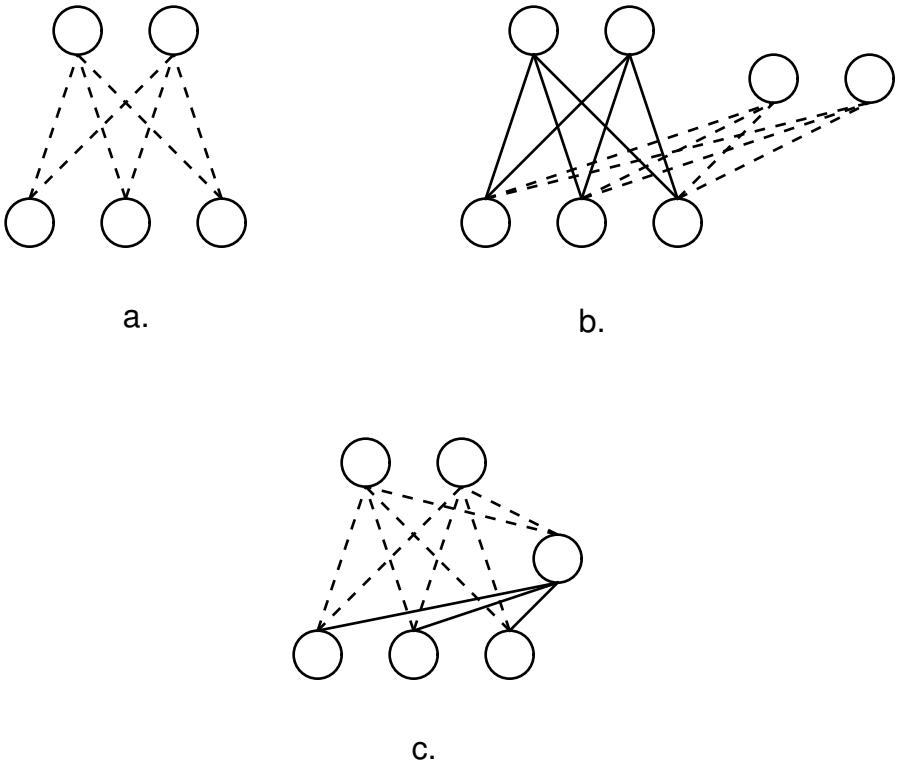


Figure 2. Hypothetical cascade-correlation nets. Modifiable weights are represented by dashed lines, frozen weights by solid lines. Output phases are shown in a and c. An input phase is shown in b. In each net, input units are at the bottom, output units at the top, and hidden units in the middle. The hidden units in b are candidates for installation. The hidden unit in c has been installed in the net.

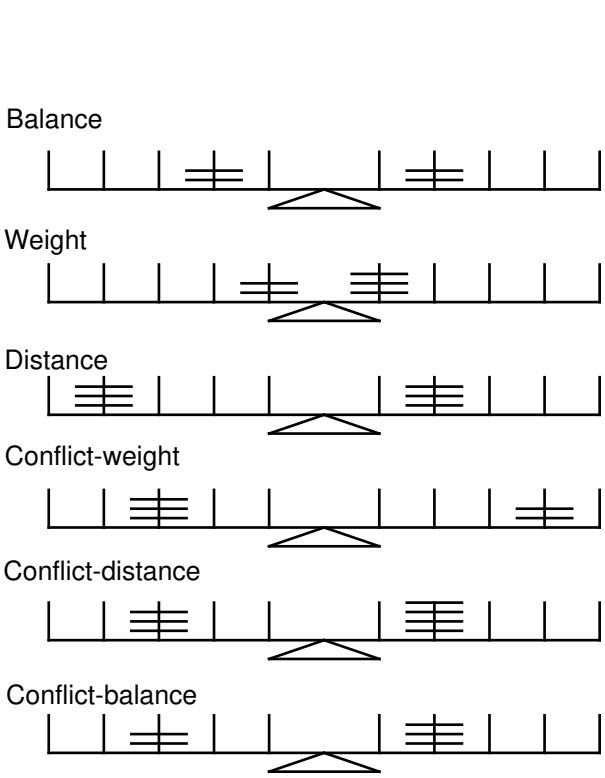


Figure 1. Sample balance scale problems and predicted success.

Table 1. Proportions of correct balance scale predictions by the addition and critical dimension rules

Problem size	Torque difference	Rule	
		Addition	Critical dimension
2	0	1.00	0.67
	1-3	1.00	1.00
3	0	1.00	0.60
	1	0.75	0.88
	2-8	1.00	1.00
4	0	0.88	0.50
	1	0.71	0.86
	2	0.78	0.89
	3-15	1.00	1.00
5	0	0.92	0.51
	1	0.64	0.75
	2	0.73	0.87
	3	0.83	0.91
	4	0.92	0.96
	5-24	1.00	1.00
6	0	0.77	0.42
	1	0.65	0.76
	2	0.68	0.80
	3	0.74	0.85
	4	0.82	0.91
	5	1.00	1.00
	6	0.92	0.96
	7-35	1.00	1.00

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. *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society*, pp. 678-685. Hillsdale, NJ: Lawrence Erlbaum.

Schmidt, W. C., & Shultz, T. R. (1991). A replication and extension of McClelland's balance scale model. Technical Report No. 91-10-18, McGill Cognitive Science Centre, McGill University, Montréal.

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

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., & Schmidt, W. C. (1991). A Cascade-Correlation model of balance scale phenomena. *Proceedings of the Thirteenth Annual Conference of the Cognitive Science Society*, pp. 635-640. Hillsdale, NJ: Lawrence Erlbaum.

Shultz, T. R., Schmidt, W. C., Buckingham, D., & Mareschal, D. (In press). Modeling cognitive development with a generative connectionist algorithm. In T. Simon & G. Halford (Eds.), *Developing cognitive competence: New approaches to process modeling*. Hillsdale, NJ: Erlbaum.

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. (1991). *Children's thinking*, 2nd edition. Englewood Cliffs, NJ: Prentice-Hall.

Siegler, R. S., & Jenkins, E. (1989). *How children discover new strategies*. Hillsdale, NJ: Erlbaum.

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

Squire, L. (1987). *Memory and brain*. Oxford: Oxford University Press.

Strauss, S. (1982). *U-shaped behavioral growth*. New York: Academic 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.

Dudai, Y. (1989). *The neurobiology of memory: Concepts, findings, and trends*. Oxford: Oxford University Press.

Elman, J. (1991). Incremental learning, or the importance of starting small. Technical Report 9101, Center for Research in Language, University of California, San Diego, CA.

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review*, **87**, 215-251.

Fahlman, S. E. (1988). Faster-learning variations on back-propagation: An empirical study. In *Proceedings of the 1988 Connectionist Models Summer School*. San Mateo, CA: Morgan Kaufman.

Fahlman, S. E., & Lebiere, C. (1990). The Cascade-Correlation learning architecture. Technical Report, CMU-CS-90-100, School of Computer Science, Carnegie-Mellon University.

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

Harnad, S., Hanson, S. J., & Lubin, J. (1991). Categorical perception and the evolution of supervised learning in neural nets. In D. W. Powers & L. Reeker (Eds.), *Working papers of the AAAI spring symposium on machine learning of natural language and ontology*, pp. 65-74.

Karmiloff-Smith, A. (1984). Children's problem solving. In M. E. Lamb, A. L. Brown, & B. Rogoff (Eds.), *Advances in developmental psychology*, Vol. 3, pp. 39-90. Hillsdale, NJ: Erlbaum.

Karmiloff-Smith, A. (1992). Nature, nurture, and PDP: Preposterous developmental postulates? *Connection Science*, *4*, 253-269.

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.

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.

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

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

Newell, A. (1990). *Unified theories of cognition*. Cambridge, MA: Harvard University Press.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*, 231-259.

Nolfi, S., Elman, J. L., & Parisi, D. (1990). Learning and evolution in neural networks. Technical Report 9019, Center for Research in Language, University of California at San Diego.

Oden, G. C. (1987). Concept, knowledge, and thought. *Annual Review of Psychology*, *38*, 203-227.

Piaget, J. (1963). *The origins of intelligence in children* (M. Cook, Trans.). New York: Norton. (Original French version published 1936)

from environmental feedback is the primary transition mechanism, producing the various qualitative and quantitative changes one sees in cognitive development.

Acknowledgements

This research was supported by a grant from the Natural Sciences and Engineering Research Council of Canada to Shultz and from the Fonds pour la Formation de Chercheurs et l'Aide à la Recherche au Québec to Mareschal.

Notes

1. Self-modifying production systems do yield qualitative changes in processing, but as will be noted later, have trouble capturing perceptual effects.

2. This formulation of C is based on Lisp code in release 11 May 1991 of cascade-correlation.

3. A preliminary report of this simulation was given in Shultz and Schmidt (1991).

4. The five peg, five weight version of the balance scale task was also used in McClelland's (1989) simulation.

5. Since the weights in these diagrams are standardized for each epoch, they may appear to vary slightly across epochs even when frozen.

6. A preliminary report of this simulation was given in Shultz and Schmidt (1991).

7. The critical dimension rule would do considerably better than it does here if it were written to predict more balanced outcomes, e.g., predict balance when the two dimensions are equally critical.

8. Note that the torque difference effect does not depend on integer coding since it is robust against considerable variation in input codes; it emerges even with codes that initially convey no direct dimensional information. In the latter cases, the network constructs quantitative dimensions during learning.

References

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: Blackwell.

Belew, R. (1993). Interposing an ontogenetic model between genetic algorithms and neural networks. In C. L. Giles, S. J. Hanson, & J. D. Cowan (Eds.), *Proceedings of the Neural Information Processing Society 5*. San Mateo, CA: Morgan Kaufman.

Brainerd, C. J. (1973). Judgments and explanations as criteria for the presence of cognitive structures. *Psychological Bulletin*, *79*, 172-179.

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.

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.

through stages 1 and 2. Although, as noted earlier, this is not the only way to produce stages 1 and 2, it is clearly one effective way to do so. Consequently, the idea of a biased learning environment can be regarded as a prediction to be tested with children. This would entail an unusual type of balance scale study that examines the natural environments of children for opportunities to observe or interact with weights and fulcrum distances. A focus on the sheer frequencies of such instances may need to be tempered with concern for what aspects of the phenomena the child is attending to.

9.6. Rule diagnosis

Our simulations uncovered some ambiguity as to whether rule 2 or rule 3 should be given precedence in the many cases in which they both fit the obtained pattern of correct and incorrect balance scale predictions. Although our principal results were relatively robust against modifications in diagnostic methods, this scoring ambiguity should probably be addressed in future psychological research on the balance scale. This is an example of how relatively precise computational modeling can help to frame empirical psychological issues.

9.7. Stage 4 performance

In our network models, reaching stage 4 is essentially a matter of learning the balance scale problem to a sufficient depth. Given sufficient experience, most of our nets eventually reach stage 4. It is quite possible that the same is true of people. Perhaps extensive balance scale experience, with corrective feedback, would lead stage 3 individuals to perform at stage 4, even without teaching them the torque rule.

10. Conclusion

The simulations reported here suggest that cascade-correlation networks can capture the main features of cognitive development on the balance scale. Combined with other simulations (Shultz et al., in press), this suggests that cascade-correlation is a particularly promising tool for modeling cognitive development. Such successful models could well provoke new theories of cognitive development, including explanations of performance on the balance scale. Much further computational and psychological work will be required to formulate any such theory in full, but we close with some speculations about the broad outline of such a theory.

A theory inspired by cascade-correlation models would view the child as being equipped with powerful, general purpose learning techniques, based primarily on pattern association, but capable of constructing new representations with greater computational power when necessary. Knowledge is represented by distributed patterns of activation across simple processing units, rather than by explicit symbolic rules. This knowledge is implicit rather than explicit, and graded rather than all-or-none. Processing occurs according to basic principles of neuronal function, such as excitation and inhibition, rather than by the matching and firing of rules. There is a tight integration of perceptual and cognitive processes, as opposed to an artificial separation between them. Such a system learns from environmental feedback, based primarily on correlations among events and is sensitive to biases afforded by the training environment. Qualitatively new representational skills emerge as required to reduce the discrepancy between expectations and results. When these qualitatively new structures do emerge, they operate on the output of existing structures when appropriate, thus ensuring a gradual hierarchical development that utilizes, rather than ignores, earlier contributions. Evolutionary pressures may affect the topology, processing and learning mechanisms, parameters, and even the initial weights in this type of system. Learning

the network's sensitivity to frequencies in the training patterns. For example, the presence of many different regular forms temporarily interferes with performance on smaller numbers of irregular forms (Plunkett & Marchman, 1991).

The present model exhibited two U-shaped effects. One instance occurred in stage 1, where error on the total set of testing problems increased in response to an increased focus on weight information (Figure 9). Although it is already well-known that stage 1 children get fewer balance scale problems correct than children in higher stages, the novelty of this prediction lies in a comparison to children before stage 1. These very young children may make more correct predictions overall than their stage 1 counterparts, a fairly counter-intuitive prediction. The explanation for this U-shaped effect would lie in the sensitivity to differential frequencies of items in the training set.

The second U-shaped effect in our nets occurred just after hidden unit recruitment. As noted in Figure 9, there were often sharp discontinuities in error reduction at these points, usually an increase in error followed by a decrease. This U-shaped effect can be attributed to the need for additional adjustment of output-side weights after the installation of a new hidden unit. In longitudinal studies of children, independently measured representational changes may accompany increases in error. If this prediction could be verified, then representational change could perhaps be indexed by an abrupt increase and decrease in error.

This U-shaped effect can be compared to Karmiloff-Smith's (1984) three phase model of children's problem solving. She described phase 1 as an error-decreasing adaptation to external information resulting in successful implicit procedures. In phase 2, children simplify their procedures into an implicit "theory-in-action" that ignores some of the precise adaptation seen in phase 1, thus increasing error. Phase 3 is characterized by an integration of the adaptational and re-descriptive accomplishments of the first two phases, resulting in error reduction and richer representations. The transition between Karmiloff-Smith's phases 1 and 2 appears somewhat analogous to the increase in error observed in cascade-correlation nets just after the transition from input phase back to output phase following the installation of a hidden unit that represents the network's output in a novel way. In both cases, new representations temporarily interfere with previous performance. The transition to Karmiloff-Smith's phase 3 is likewise analogous to cascade-correlation's eventual adjustment of output-side weights after hidden unit installation. Error decreases and the network's representation of the problem being learned is better tuned than ever.

A major difference between the two approaches is that the re-description in Karmiloff-Smith's phases 2 and 3 is driven by an as yet unspecified analysis of earlier procedures and theories, not by performance error. In contrast, both the input and output phases in cascade-correlation are driven by the necessity to reduce error. The output phase reduces error directly by adjusting output-side weights; and the input phase creates a hidden unit whose activations re-describe the network's output in a way that correlates with existing error. Another difference is that cascade-correlation may not create the explicit awareness of knowledge that is often credited to children, for example in Karmiloff-Smith's (1984, 1992) phase 3. At this point, it is very unclear how or whether connectionist models can model such a transition. Nonetheless, the present simulations show that at least some developmental increases and decreases in error can be accounted for by a single, homogeneous error-correction mechanism.

9.5. Biased learning environment

A major assumption underlying both our model and McClelland's (1989) model is that a learning environment biased in favor of equal distance balance scale problems causes networks to pass

One of the model's chief predictions, the torque difference effect, is already an established psychological result. We insist on calling it a prediction because we discovered it in cascade-correlation nets before any of us knew of the Ferretti and Butterfield (1986) paper. That none of the several previous computational papers on the balance scale have even mentioned the torque difference effect attests to the fact that it is not nearly as well known as the stage progressions.

Our networks suggest that the torque difference effect is related to the perceptible magnitude of the information entering the computational system. A problem is easier to solve if its torque difference is larger. The torque difference effect occurs in connectionist networks because the magnitudes of the inputs reflect the numbers of the weights and the magnitudes of the distances in the problem. This suggests that exaggeration of weight and distance quantities could enhance the torque difference effect in children. The locus of the torque difference effect could also be focused on a particular dimension, by making either the numbers of weights or the magnitudes of distances larger. Again these effects could not be captured by rules that were sensitive only to difference directions, but not to magnitudes.

Figure 9 reports that the lowering of error over learning time increases with torque difference. Note also in this figure that there is a spreading out of the different torque levels as learning progresses. This implies that, with increasing age, the torque difference effect would become more pronounced. This makes sense since the notion of torque requires integration of both distance and weight information. According to the error plots in Figure 9, the lack of a torque difference effect should be most pronounced prior to, or early in, stage 1. This would presumably be a time of mostly random predictions, even on simple balance scale problems.

9.2. Tentative transitions with non-random behavior

The cascade-correlation model suggests that balance scale stage transitions are soft and tentative rather than abrupt and definitive. However, during these transition periods, the nets continue to have diagnosable behavior that alternates between two successive stages over proximal testing sessions. This suggests that, during stage transitions, children would continue to show distinct, classifiable behaviors rather than perfoming in a random, uninterpretable fashion. Testing of these predictions would, of course, require a longitudinal research design.

9.3. Stage skipping and regression

Stage skipping and regression to earlier stages did occur in our networks. Longitudinal balance scale studies of children, with repeated testing on small time slices, could be used to investigate these network predictions more thoroughly than has been possible in existing cross-sectional designs. Again, such effects would be difficult to capture in symbolic rule-based models. In most such models, once a behavior has become prominent it remains so until superseded by the next stage.

9.4. U-shaped development

U-shaped development has been interesting to psychologists because it deviates from the general, and more obvious, trend for children to improve their performance with age and practice. Numerous examples of U-shaped development have been documented in the developmental literature (Strauss, 1982), but coherent theoretical explanations of it have been lacking. Past modeling has shown that U-shaped development can occur as a natural result of connectionist learning in certain environments. Generally, U-shaped development can be understood in terms of

balance scale problem. In contrast, children rarely think solely about balance scales or weights for long periods of time because they are more fully engaged with a complex ongoing flux of other, usually more pressing, problems.

Nonetheless, along with other connectionist simulations, this work suggests that a connectionist approach can successfully model aspects of children's cognitive development. We are particularly keen on using generative algorithms, such as cascade-correlation, for this modeling. With a generative algorithm, not only is network design more principled, but the structure of the network undergoes qualitative increases in representational power. Such changes may well underly some observed stage transitions.

It is also possible to achieve large qualitative changes in behavior through continuous small changes in the weights inside a network of fixed topology. Such outcomes can be described in terms of mathematical catastrophe theory (Pollack, 1990; van der Maas & Molenaar, 1992). That is, qualitative behavioral changes can arise either from a major restructuring of cognitive processing or from continuous small changes of chaotic systems. The former is the traditional view of cognitive development (Piaget, *passim*), whereas the latter is a major principle in catastrophe theory. A possible advantage of cascade-correlation models is that both types of transition mechanisms can be examined simultaneously. Results so far suggest that some transitions occur when hidden units are recruited, whereas others occur through weight modifications. The fact that cascade-correlation nets sometimes fit psychological data better than do topologically static nets suggests that both transition mechanisms are required. Alternate systems based exclusively on either quantitative (e.g., McClelland, 1989) or qualitative (e.g., Newell, 1990) transition mechanisms do not facilitate study of the interaction among qualitative and quantitative processes.

The qualitative-quantitative distinction can also be viewed in terms of Piaget's (1936/1963) distinction between assimilation and accommodation. Assimilation is the child's tendency to distort incoming information so that it better fits the child's existing cognitive structures. Accommodation is the contrasting tendency to adjust internal cognitive structures so that they more accurately reflect incoming information. Together assimilation and accommodation describe, albeit vaguely, the child's adaptation to the external world. These adaptational processes can be re-interpreted in terms of the computational mechanisms of the cascade-correlation algorithm. Accommodation can be understood as the recruitment of hidden units, the building of qualitatively different representational power. Assimilation can be viewed as correct generalization to novel cognitive tasks, requiring neither hidden unit recruitment nor weight changes. Cascade-correlation also allows a concrete interpretation of assimilative learning, something that Piaget never really explained. Assimilative learning can be understood as the quantitative adjustment of weights in a network, without the necessity of undergoing qualitative topological change. This represents not only a novel interpretation of Piaget's theory of cognitive adaptation, but an important extension to it.

9. Predictions and future work

One of the most useful features of a detailed computational model is the ability to generate predictions for new psychological research. Our cascade-correlation balance scale model suggests a number of different avenues to explore, within the context of both further computational modeling and empirical investigation of children's behavior.

9.1. Torque difference

We have found that a more "nativist" simulation in which the network starts (by virtue of some pre-training) with knowledge of how weight information affects balance scale performance also captures the correct stage progression, even without biased training patterns (Shultz, Schmidt, Buckingham, & Mareschal, in press). These networks were initially trained only on weight problems (Figure 1). Then the training set was changed to include all six problem types in an unbiased fashion. In addition to correct stage progressions, these nets also showed the torque difference effect. It is critical for capturing correct stage progressions that the network emphasize weight information at some point early in its history, but it appears that this emphasis can be achieved in more than one way. This knowledge can either be wired in, or it can reflect bias in the learning environment, or presumably some combination of both. Further psychological and computational work is obviously necessary to determine the best model of the child's cognitive development on the balance scale. Examination of behavioral subtleties not yet available in the psychological literature may be necessary to sort out the various candidate models. At this early stage, it is important to note that connectionist techniques are not bound exclusively to an experiential approach to the neglect of nativistic concerns. Indeed, connectionist networks seem ideally suited to investigate the integration and interaction of experiential and innate forces (Bates & Elman, 1993; Belew, 1993; Karmiloff-Smith, 1992; Nolfi, Elman, & Parisi, 1990).

Like other connectionist models of developmental phenomena, the present simulations suggest that the connectionist approach deserves serious consideration as a means of studying transition mechanisms for higher level reasoning. Connectionist networks appear capable of reproducing classic developmental phenomena such as rules and stages, as well as more subtle effects such as information salience that explicit symbolic rule systems have particular difficulty with.

An explicit symbolic rule-based model trying to capture the torque difference effect would presumably find itself in the paradoxical position of having to compute torque differences well before stage 4. It might require rules of the form *if torque difference is greater than x then apply rule i*, where *x* is some integer between 1 and 20 that decreases with age, and *i* is the current stage. Such a model would apparently have to compute and use torque differences to mimic the torque difference effect well before it could compute and use torques to solve balance scale problems. This would possibly fit the psychological data, but would be extraordinarily awkward.

The reason that connectionist network models are able to capture the torque difference effect so naturally is based on their sensitivity to the amounts of weight and distance. Networks such as cascade-correlation are naturally sensitive to input magnitudes. Activations of the units in the network are a continuous function of the magnitudes of the inputs. The hidden and output units are relatively more affected by inputs of greater intensity. Thus, larger and more distinctive differences in weight or distance inputs will tend to yield clearer activation patterns on the hidden units and consequently more decisive predictions on the output units. In contrast to this, symbolic rules encode and combine discrete values of weight and distance. Rules are thus sensitive only to the direction of weight and distance differences, but not to the extent of these differences. In summary, computation is continuous in networks, but discrete in symbolic rules; continuous computation would appear to be essential for phenomena like the torque difference effect.[8]

The present model is, like other current models, highly simplified compared to the actual learning environment and computational resources of children. There are two principal differences between children and these networks, one of them favoring the children and the other favoring the networks. First, children have the advantage of applying their considerable knowledge to the balance scale task, whereas the networks start learning about the balance scale from scratch, that is, from a set of random weights. The interaction of network learning with pre-existing network knowledge has not yet been much studied, but is likely to attract considerable attention in the near future. Second, networks have the advantage of being able to devote their full resources to the

These proportions correct are presented in Table 1. At most levels of problem size and torque difference, both rules generated perfectly correct predictions, i.e., proportions of 1.00.[7] It is only at very low torque differences that these two rules falter a bit.

==========Table 1==========

7.3. Discussion

These results do not support the claim that use of these two rules can capture the torque difference effect. This is because, over a wide range of torque differences, the rules fail to vary in their rate of success. The research with children showed variation in success over a wide range of torque differences (Ferretti & Butterfield, 1986) as did the cascade-correlation nets reported in simulation 3 (section 6).

To put this another way, there is a need to produce the torque difference effect at all four balance scale stages. The overall high rate of correct performance generated by these two rules suggests that they would function mainly at the higher stages. But the results with children show the torque difference effect at every stage of development, even in stage 1. To produce torque difference effects in stage 1 with symbolic rules would require very simple rules indeed to compete with the ultra-simple weight rule that defines stage 1. Such rules would have to be far less successful that the rules simulated here to yield the low rate of success characteristic of stage 1.

In contrast to these apparent difficulties with a symbolic rule approach, cascade-correlation nets capture the torque difference effect at every stage, with no ad-hoc assumptions. The torque difference effect falls naturally out of the network solutions.

8. General discussion

In these simulations, cascade-correlation networks learned to perform on balance scale problems as if they were following rules, including clear performance at the level of the rule that characterizes stage 4. Further, these stages tended to emerge in the psychologically correct order. Some developmental regressions and stage skipping were observed and the transitions between stages tended to be soft and tentative. Psychological longitudinal studies suggest that all of these phenomena are characteristic of cognitive development in children (Siegler & Jenkins, 1989). The cross-sectional research designs used with children on the balance scale are less than ideal for investigating issues of regression and skipping. Stage skipping, in particular, would require very small time slices to be sure that children actually missed a stage. Some regression to earlier balance scale stages, however, has been noted in existing cross-sectional research (Chletsos, De Lisi, Turner, & McGillicuddy-De Lisi, 1989; Siegler, 1981).

Unlike previous models, these nets also captured the torque difference effect. We had actually predicted the torque difference effect from early simulation results before discovering the Ferretti and Butterfield (1986) paper in the literature.

The cascade-correlation networks covered the relevant psychological phenomena without at least some of the restrictive assumptions of McClelland's (1989) back-propagation networks. We didn't need to implant segregated hidden units for weight vs. distance information. Indeed, because we were using cascade-correlation, we did not initially implant any hidden units at all. We did, however, follow McClelland's lead in strongly biasing the training patterns in favor of equal distance problems. Such input bias is not the only way to obtain human-like stages in connectionist models of the balance scale, but it's effectiveness in producing stages may encourage researchers to examine biases in the child's learning environment.

instances of U-shaped development. This is seen in Figure 9 as an inverted U since error is being plotted, rather than correct performance.

To assess the torque difference effect more systematically, an ANOVA of these error signals midway (epoch 75) and late (last epoch) in learning was performed across the 16 networks. In this analysis, torque difference level and epoch served as within net factors. This analysis yielded only a main effect for torque difference level, $F(3, 42) = 48.57$, $p < .001$, with a strong negative linear trend, $F(1, 42) = 140.45$, $p < .001$. The mean errors at these two epochs for the four torque difference levels are presented in Figure 10. In general, the larger the torque difference, the smaller the error as could have been predicted from the psychological data. Thus, cascade-correlation networks easily capture the torque difference effect that has been observed with children.

==========Figure 10==========

7. Simulation 4: Symbolic rules for torque difference?

In describing these torque difference results to a leading balance scale researcher and noting the apparent difficulty that rule-based models have with this phenomenon, one of us elicited the counter-argument that there could well be some simple, alternate rules that children might follow in conformity with the torque difference effect. That is, even though the rules embodied in Siegler's four balance scale stages cannot seem to capture the torque difference effect, perhaps other simple rules could.

We investigated two such suggested rules. One could be called the *addition* rule. It specifies that the side with the larger sum of weight + distance will go down. When the sums on the two sides are equal, presumably this rule would predict that the scale should balance.

The other rule could be called the *critical dimension* rule. It first defines the critical dimension, either weight or distance, as the dimension that produces the larger absolute difference. That is, which is larger: absolute (left weight - right weight) or absolute (left distance - right distance)? Then the rule specifies that the side with the larger value on that critical dimension will be the one to drop. When the two absolute differences are equal, we can specify that the weight dimension is favored. Under such circumstances, if it is furthermore true that the weights are equal, we can predict that the scale should balance.

If either of these two rules could produce the torque difference effect, this would contradict our argument that only connectionist models, and not symbolic rule-based models, could capture this effect.

7.1. Method

We tested these two rules with a simple simulation. We first generated all possible balance scale problems of sizes 2, 3, 4, 5, and 6. Size in this context refers to the maximum number of weights and also to the maximum number of pegs on each side of the fulcrum. There was the usual restriction that all weights on each side of the fulcrum are placed on a single peg. Then we computed the proportion of correct balance scale predictions given by the addition and the critical dimension rules for each size of problem and each torque difference.

7.2. Results

The finding of higher stages with more learning in cascade-correlation nets is relatively robust against these variations in scoring method. To our knowledge, there has been no analogous comparison of these scoring variations in studies with children. Indeed, the diagnostic ambiguity arising from the fact that many behavior patterns can equally well be classified as rule 2 or rule 3 has not been acknowledged in psychological research. Nor is it clear how these diagnostic ambiguities may have been resolved in various studies. As noted in section 4.4, we favor giving scoring priority to rule 2 over rule 3 because rule 2 is characterized by a definite advantage in number correct on conflict-weight problems (Figure 1). The drop in performance on conflict-weight problems is an instance of U-shaped development, wherein children get worse before they get better. Although it is true that stage 3 performance is supposed to be better than stage 2 performance on conflict-distance and conflict-balance problems, this is attributed to a *muddle through* strategy, whereby children simply guess on conflict problems (Figure 1; Siegler, 1976, 1981). The more definite superiority in performance between stages 2 and 3 is provided by stage 2 children performing at near perfect level on the conflict-weight problems as compared to the chance-like muddling through of stage 3 children on those problems. So, on the grounds of giving priority to systematically correct performance, stage 2 should have scoring priority over stage 3.

6. Simulation 3: Torque difference effect[6]

The purpose of this simulation was to test for the torque difference effect. This is the idea that problems with large torque differences are easier to solve than those with small torque differences.

6.1. Method

Exactly the same networks and techniques were employed as in the first two simulations except that the principal interest was in recording errors at four different torque difference levels: 1, 2-5, 6-9, and 10-20. Consequently, the testing patterns had to be quite different for this simulation. For each of 16 runs, four sets of test patterns were randomly selected from each of Siegler's six problem types. Each set of test patterns contained only problems representing one of the four torque difference levels.

We assessed ease of solution by examining the network's error on problems of different torque difference during learning. Error was computed as the sum of squared discrepancy between actual and target outputs, as shown in Equation 2, summed over the 24 problems of each torque difference level.

6.2. Results

Errors are plotted over epochs for two representative networks in Figure 9. Only every fifth epoch is plotted to increase clarity. As expected, each network showed faster and deeper error reduction with increasing torque difference.

==========Figure 9==========

In addition, two serendipitous findings emerged from these error plots. First, the sharp discontinuities in the error plots coincide with the installation of new hidden units. Presumably, error can get a little worse until the output weights from the new hidden unit are adjusted early in the next output phase. Second, the emergence of stage 1 at around 25 epochs was typically characterized by an increase in error. That is, a strong focus on weight information worsens performance on the test problems as a whole. Both of these phenomena may be regarded as

5. Simulation 2: Diagnostic criteria and stage assessment

Simulation 1 utilized a rule diagnosis scheme that was very much in the spirit of those used by psychological researchers and other computational modelers of balance scale phenomena. It is not, however, the only available diagnostic scheme. Consequently, the purpose of this simulation was to examine the impact of certain variations in diagnostic criteria on stage assessment in cascade-correlation networks.

5.1. Method

Two principal scoring variations were studied. One involved the imposition of two additional scoring criteria introduced by Siegler (1976, 1981) for diagnosing balance scale rules in children. He required at least three correct responses to the four distance problems in the test set for a diagnosis of stage 2 or stage 3; and he required three or more incorrect responses on distance problems for a diagnosis of stage 1. We refer to this as *Siegler* scoring to distinguish it from the somewhat less rigorous *SMS* scoring system used in simulation 1.

The other scoring variation was to give priority to stage 3 over stage 2, rather than 2 over 3 as in the previous simulation. Thus, four different stage diagnosis schemes were employed here, termed *Siegler2*, *Siegler3*, *SMS2*, and *SMS3*. The numerical suffix indicates the stage that has priority in case the criteria for more than one stage are satisfied. Only stages 2 and 3 suffered from this sort of diagnostic ambiguity. Apart from these variations in rule diagnosis, the simulations proceeded exactly as those in simulation 1.

5.2. Results

Each of 16 network runs yielded a series of rule diagnoses over epochs of learning. The canonical, predicted series of 1 2 3 4 was the most frequent outcome for the Siegler2, SMS2, and SMS3 scoring methods. For the Siegler3 method, which lacked many diagnoses of rule 2, the most common pattern was 1 3 4. A good deal of regression and stage skipping can be observed within all four scoring methods.

We needed a metric with which to assess the various scoring methods for their ability to capture the psychological data on stage progression. Using the rule progression data just described, we noted the position occupied by each rule. The positions in the diagnostic series occupied by each of the four rules were subjected to an ANOVA in which the rules, with four levels, and diagnostic methods, with three levels (Siegler2, SMS2, and SMS3), served as factors. Each rule occurrence contributed a single data point in this analysis. This analysis yielded a large main effect of rule, $F(3, 177) = 44.25$, $p < .0001$, and a small main effect of method, $F(2, 177) = 3.95$, $p < .05$. There was no method by rule interaction.

Mean positions of each of the rules in the diagnostic stage series for all four diagnostic methods are presented in Figure 8. A linear contrast test for the three methods included in the ANOVA yielded $F(1, 177) = 127.56$, $p < .0001$. Each successive rule tends to occur in successively higher serial positions of the diagnostic record, regardless of the diagnostic method. Rule 1 tends to occur first, rule 2 second, rule 3 third, and rule 4 last, all of which is consistent with the cross-sectional psychological data.

==========Figure 8==========

5.3. Discussion

to the product of side and distance or side and weight, and two were mainly sensitive to the bias unit or to older hidden units.

4.7. Discussion

This first simulation showed that progression through the four rule-like stages of the balance scale can be captured by cascade-correlation networks, even though rules are not explicitly represented anywhere in the networks. Critical assumptions included a learning environment that is heavily biased towards equal distance problems and gradually exposes the network to more balance scale problems. Unlike the back-propagation network of McClelland (1989) these cascade-correlation nets did not require a pre-designed hidden layer, segregated with separate channels for weight and distance information. Also, in contrast to the back-propagation networks, the cascade-correlation nets could remain in stage 4 without sacrificing an earlier progression through stages 1 and 2.

There has been an informal debate among balance scale modelers about the advisability of capturing stage 4 performance, particularly in view of the fact that many models do not naturally end up in stage 4. Much of this debate is fueled by the notion that people who perform at stage 4 might be using an explicit torque rule (Siegler & Klahr, 1982). Although not everyone reaches stage 4 performance in psychological studies, clearly some individuals do (Siegler, 1976, 1981). Therefore, it is our view that the ability to perform at stage 4 is a critical feature of balance scale models. The fact that many stage 4 performers can justify their predictions by citing a version of the torque rule does not mean that they are explicitly computing and comparing torques. There is a close, albeit imperfect, correspondence between diagnosed stage and verbal justification at each of the four stages of the balance scale and related problems, not just at stage 4 (Siegler, 1976, 1981). Moreover, children and adults as well are not particularly accurate in reporting on their cognitive processes after the fact (Brainerd, 1973; Ericsson & Simon, 1980; Nisbett & Wilson, 1977). Data showing that stage 4 adults respond more quickly to non-conflict problems than to conflict problems (Siegler & Klahr, 1982, pp. 143-144) are likewise not definitive in establishing explicit use of torques since the cause of these reaction time differences is not yet clear. Problem type and torque difference are typically confounded in psychological studies, often with conflict problems having much smaller torque differences than non-conflict problems. Either conflict or low torque difference levels could make problems more difficult. Further, it is unknown whether similar reaction time differences might also occur at earlier stages. The fact that diagnosis at stage 4 in children between 6 and 11 years varies (from 5% to 37%) with torque difference level (Ferretti & Butterfield, 1986) suggests that explicit use of the torque rule cannot be the entire story. If the torque rule were being explicitly applied, it should, like other rules, apply regardless of the amounts of weight and distance differences (see General Discussion).

It is important to realize that this simulation employed a longitudinal research design in which each network's performance on the balance scale task was assessed over time. This is in contrast to psychological studies of the balance scale, which have so far only involved cross-sectional designs in which several age groups were sampled at a particular point in time. The relative efficiency of the cross-sectional design has ensured its popularity with psychologists as compared to the time consuming and expensive longitudinal design. Longitudinal designs, however, are capable of providing much richer detail on developmental changes. It is unclear whether longitudinal research designs applied to children would yield the orderly progressions found in this simulation. In this respect, the longitudinal method applied here to networks may be a bit too rigorous and conservative for the psychological data. The longitudinal method was favored for the simulations, however, because it provided a finer grain of analysis than the cross-sectional psychological studies were able to, and because a longitudinal design is not difficult to execute with simulations.

To better understand developing network structure and the role of hidden units, we drew Hinton diagrams in the middle of each rule-based stage. Each such diagram shows the size and sign of incoming weights at a particular epoch. Hinton diagrams for two representative networks are presented in Figures 6 and 7. Each gray strip in a diagram contains the weights coming into a hidden or output unit from the various sending units that are numbered across the top of the strip. The size of each weight is indicated by the size of a corresponding square; the sign of each weight is indicated by the color of the corresponding square, with white indicating positive and black negative. The precise output training epoch from which the weights were taken is also indicated for each Hinton diagram.[5]

==========Figure 6==========

Figure 6 shows Hinton diagrams for a net that adds a single hidden unit. During stage 1, in which children use only weight information, the output units were highly sensitive to weight information coming in from input units 2 and 4. The right-side down output received an excitatory signal from the right-side weight input (unit 4) and an inhibitory signal from the left-side weight input (unit 2). The opposite was true for the left-side down output unit: it received an excitatory signal from the left-side weight input (unit 2) and an inhibitory signal from the right-side weight input (unit 4). In the midst of stage 2, in which children continue to use weight but begin to use distance information when the weights on each side are equal, the network's outputs became more sensitive to distance information. The differential sensitivity to sides was retained. In the middle of stage 3, which is characterized by children's use of both weight and distance information but confusion when these are in conflict, the output units continued to become more sensitive to distance information. More importantly, a new hidden unit had been added that precipitated the jump to stage 3. This hidden unit was particularly sensitive to side information, with strong excitatory signals from right-side inputs and strong inhibitory signals from left-side inputs. During stage 3, this hidden unit developed an excitatory link to the right-side down output and an inhibitory link to the left-side down output. Finally, in stage 4, which signifies nearly correct performance, there was additional fine tuning of these connections, but no major qualitative shift.

A different solution is portrayed in Figure 7 for a net that recruited two hidden units. During stage 1 (weight information only), the output units were highly sensitive to weight information, much as in the net featured in Figure 6. Again, the right-side down output received an excitatory signal from the right-side weight input (unit 4) and an inhibitory signal from the left-side weight input (unit 2). The left-side down output received an excitatory signal from the left-side weight input (unit 2) and an inhibitory signal from the right-side weight input (unit 4). In stage 2 (mainly weight information, but distance information when the weights on each side are equal), the network's outputs became far more sensitive to distance information. The differential sensitivity to sides was retained, and the new hidden unit was particularly sensitive to weight information (from units 2 and 4). In stage 3 (use of both weight and distance information but confusion when these are in conflict), the outputs became just about as sensitive to distance as to weight. Finally, in stage 4 (nearly correct performance), a new hidden unit emerged that was particularly sensitive to distance information. The two recruited hidden units, the first representing mainly weight and the second mainly distance, sent essentially opposite signals to the output units. The first hidden unit (mainly sensitive to weight) inhibits the right-side down output and excites the left-side down output. The second hidden unit (mainly sensitive to distance) does the opposite although less strongly.

==========Figure 7==========

More generally across the 16 networks, we found that, of the 21 hidden units with relevance to the Hinton diagrams, eight were especially sensitive to side information, 11 were mainly sensitive

four patterns from each of Siegler's six problem types (see Figure 1). For each problem type, one pattern was selected from each of four levels of torque difference: 1, 2-5, 6-9, and 10-20. Before each epoch during the output phase, the network was tested with these 24 test patterns. Any test problem in which both resulting output activations were within score-threshold of their correct targets was coded as correct; any other test problems were coded as incorrect. Past simulations (McClelland, 1989) and psychological research (Siegler, 1976, 1981) have confounded torque difference level with problem type, thus making the diagnosis of rules somewhat ambiguous. Our balanced test patterns eliminate this confound.

The patterns of correct and incorrect problems were used to diagnose rule use, in the spirit of Siegler's (1976, 1981) methods with children. A diagnosis of stage 4 required 20 or more of the 24 test problems correct. Stage 2 required 13 or more correct on the 16 balance, weight, distance, and conflict-weight problems and less than three correct on the eight conflict-distance and conflict-balance problems. Stage 3 required ten or more correct on the 12 balance, weight, and distance problems and less than ten correct on the 12 conflict problems. Stage 1 required 10 or more correct on the 12 balance, weight, and conflict-weight problems and less than three correct on the 12 distance, conflict-distance, and conflict-balance problems. The priority of scoring for these stages, in decreasing order, was 4, 2, 3, and 1. Stage 2 was given a higher priority than stage 3, because rule 2 produces fewer errors on conflict-weight problems, as shown in Figure 1.

4.5. Results: stage progressions

Figure 4 plots the stage diagnosed at each output training epoch for a representative network. The symbol *H* on the bottom of the plot signifies the epochs at which hidden units were added to the network.

==========Figure 4==========

Tabulation of rule diagnosis results across all 16 networks revealed that 11 showed the predicted 1 2 3 4 ordering. Two other nets showed rules 1 2 3; one showed rules 1 2 4; one showed rules 1 2 4 with regression to 3 and 2; and one showed rules 1 2. With continued training beyond 300 epochs, such nets do tend to converge on stage 4.

The overlap between diagnoses of adjacent rules near transition points in Figure 4 reflects the tentative nature of some transitions. There is often a period of going back and forth between two stages before settling on the higher level stage.

Of the 16 nets, nine of them recruited a single hidden unit, six recruited two hidden units, and one recruited three hidden units. Of these 24 hidden units, 13 were associated with a quick progression from one stage to the next: five advanced to stage 4, seven to stage 3, and one to stage 2. The other nine hidden units may have played a role in maintaining the current stage or in more gradually preparing the way for the next stage, but this is more difficult to verify. The net whose results are displayed in Figure 4 appeared to utilize its two hidden units to progress to stages 3 and 4, respectively.

Mean stages across all 16 nets are plotted in Figure 5 over the first 200 output training epochs. Although the subtleties of individual network plots are obscured by the averaging, Figure 5 does reveal a clear increase in stage performance as learning progresses.

==========Figure 5==========

4.6. Results: network analysis

hidden unit to the second hidden unit. Both networks 3a and 3b are drawn with activations reflecting the input and output coding of the balance scale problem shown in Figure 3c.

4.2. Training

There were 100 initial training patterns. They were randomly selected without replacement from the 625 possible five peg, five weight problems, subject to a 0.9 bias in favor of equal distance problems (*balance* and *weight* problems, as illustrated in Figure 1). This selection bias ensured that the probability of drawing an equal distance problem during construction of the training patterns was 0.9. On each epoch in the output phase, another training pattern was randomly selected with replacement, also subject to the 0.9 equal distance bias, and added to the training patterns. We refer to this as *expansion* training of the *1+* type. The training set is gradually expanded, with one new pattern added on each epoch of the output phase. Expansion training conforms to our assumptions that the child's learning environment changes gradually and that these changes are marked by exposure to more aspects of the environment. This is in contrast to the training regime in McClelland's (1989) model which used a completely fresh random selection of training patterns each epoch. The constant bias in favor of equal distance problems reflects the assumption, originally made by McClelland (1989), that children have plenty of experience lifting differing numbers of objects but relatively little experience placing objects at discrete different distances from a fulcrum.

Our pilot simulations had established that, without the strong bias for equal distance problems, the network skipped stages 1 and 2, proceeding directly to stages 3 and 4. Although bias did not need to be constant across epochs to produce correct stage progressions, constant bias was considered to be a simpler assumption than a changing environment. Other pilot simulations indicated that learning was extremely difficult when 100 training patterns were randomly selected each epoch. Still others showed that, without expansion training, the network would learn the training patterns before being able to progress very far with the stages, as assessed using the testing patterns (see section 4.4).

4.3. Parameters

We used default parameter values for cascade-correlation (Fahlman & Lebiere, 1990), with two exceptions. We lowered the input and output Epsilons (learning rates) by 1/2 in order to reduce the bounce in errors from epoch to epoch. Also, we used a score-threshold of 0.25, rather than the default value of 0.4 which is appropriate for sigmoid units coding dichotomous target values. Because our output units were also coding neutral (balance) patterns, we lowered score-threshold to 0.25 in order to achieve non-overlapping scoring ranges. An output activation had to be equal to or greater than absolute 0.25 to count as a tipped balance beam.

Each simulation had 16 networks or runs. Each run was terminated at 300 epochs because pilot simulations had established that most runs were well within stage 4 by that epoch. Networks did not typically declare victory on the training patterns within 300 epochs because expansion training kept the network a bit off balance; each new epoch a possibly novel pattern was added to the training patterns.

4.4. Diagnosing rule use

Each of the 16 runs used distinct, randomly selected training and test patterns. The 24 test patterns in this simulation were balanced for both problem type and torque difference, so that there were

During the input training phase, the weights leading to the output units are frozen. 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 so as to maximize the absolute value of a modified correlation (*C*) between the activation of that unit and the residual error at the output units, across all pattern pairs:

$$C = \frac{\sum_o \left| \sum_p \left(h_p - \langle h \rangle \right) \left(e_{op} - \langle e_o \rangle \right) \right|}{\sum_o \sum_p \left(e_{op} - \langle e_o \rangle \right)^2} \tag{3}$$

where h_p is the activation of the candidate hidden unit for pattern *p*, <*h*> 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.[2]

The input training phase continues until *C* stagnates or until a specified maximum number of epochs has elapsed. At this point, the candidate unit with the largest *C* is retained while all the other candidate hidden units are discarded. The input weights to the new hidden unit are then permanently frozen and the new hidden unit is connected to all of the output units. The algorithm then returns to the output training phase with the added power of a new unit that is particularly good at detecting the network's current residual error.

Because candidate hidden units receive connections from all input and current hidden units during the input phase, any network topology can in theory be constructed by the cascade-correlation algorithm. It might be assumed that, since each new hidden unit is installed on its own layer, it would not be possible to achieve a back-propagation style net with multiple hidden units within a single hidden layer. However, such an arrangement can be achieved in cascade-correlation if cascaded hidden-to-hidden weights become 0 during training.

4. Simulation 1: Balance scale stages[3]

4.1. Network design

All of the cascade-correlation simulations reported here employ the same sort of network and a five peg, five weight version of the balance scale.[4] As illustrated in Figure 3a, the initial network had four input units, the obligatory bias unit (which always has an input of 1 in cascade-correlation), and two output units. Of the four input units, one encoded left-side weight, a second encoded left-side distance, a third encoded right-side weight, and a fourth encoded right-side distance. The input coding of weight and distance information was done using integers from 1 to 5. On the output side, there were two sigmoid units that represented balance scale results in a distributed fashion. Left-side down was conveyed by excitation (0.5) of the first output and inhibition (-0.5) of the second output; right-side down was conveyed by the reverse pattern; and a balanced result was conveyed by neutral values (0) on both outputs. Any hidden units that were recruited also used a sigmoid activation function.

==========Figure 3==========

Figure 3b shows the structure of a network after two hidden units have been recruited. The connections are feed-forward only, from inputs to hiddens to outputs. In addition, the output units receive direct connections from input and hidden units, and there is a connection from the first

All learning occurs in batch mode, meaning that any weight modifications occur after a complete presentation of the training input/output pattern pairs. Such a presentation constitutes an epoch. Learning continues until the responses of the output units are each within a sufficiently small value (the score-threshold) from a desired target, for all pattern pairs. At that point, the network declares victory and stops learning.

Although batch learning is often regarded as psychologically suspect, there is actually considerable psychological (Oden, 1987) and physiological (Dudai, 1989; Squire, 1987) evidence for it. For example, the hippocampus seems to learn in batch mode for later storage in the neocortex (Dudai, 1989; Squire, 1987). The hippocampus apparently stores numerous examples and then eventually abstracts and transfers the essential information to the neocortex.

The standard contrast to batch learning is pattern learning in which weights are updated after the presentation of each individual pattern. Batch learning normally requires fewer weight updates than does pattern learning because in batch learning all patterns are considered at once. In contrast, pattern learning may require the undoing of some weight changes because not all patterns were considered. Consequently, batch learning is relatively efficient.

It is also important to note that, even in batch learning, the training patterns are processed individually. That is, for each pattern, outputs are compared to their targets independently of other patterns. Thus, the system never has to process more than one pattern at a time. It must, however, keep a running sum of the error, which is eventually used to adjust the weights.

All hidden and output units in the simulations presented in this paper possess sigmoid activation functions defined by:

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

where y_i is the resulting activation of the receiving unit indexed by i, x_j is the activation of a sending unit indexed by j, and w_{ij} is the weight connecting those two units. The cascade-correlation algorithm is also compatible with a number of other activation functions including linear and gaussian.

3.1. Output training phase

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

$$E = \sum_o \sum_p \left(A_{op} - T_{op} \right)^2 \tag{2}$$

where o indexes the output units, p indexes the input-output pattern pairs, A is the actual activation of an output unit, and T is the target activation for that output unit. If either 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 training phase.

3.2. Input training phase

3. Cascade-correlation

In contrast to the static, user-designed networks characteristic of previous developmental connectionist research, we favor the use of the cascade-correlation algorithm (Fahlman & Lebiere, 1990) for modeling cognitive development. Like other so-called generative algorithms, cascade-correlation constructs its own network topology as it learns. It does this by recruiting new hidden units into the network as it needs them to solve a problem.

This generative technique affords a more principled approach to network construction than is typical of connectionist research. Instead of merely adjusting weights in a network of fixed topology, cascade-correlation starts with a minimal network of input and output units. During learning, it may add hidden units one at a time, installing each on a separate layer. If the net is not reducing error fast enough with its current topology, it will select and install a new hidden unit whose output activations correlate best over all training cases with the existing network error. In essence, cascade-correlation searches not only weight space but also the space of network topologies.

The principal advantage of generative approaches for simulating cognitive development is the ability to model qualitative changes in representational power as well as the more gradual quantitative adjustments in network weights. Qualitative changes in cognition mean that the child comes to process information in a distinctly different way than before, not merely that the child is faster, has more stored information, or has a larger memory span. Such qualitative leaps have long been considered inherent to children's development (Piaget, *passim*), but until now have been difficult to model in a rigorous way that provides a good fit to psychological data.[1] Cascade-correlation affords a novel and natural interpretation of both qualitative and quantitative developmental changes. Qualitative changes occur through the recruitment of new hidden units, and quantitative changes through the adjustment of network weights. Psychologically, new hidden units might correspond to processing structures that transform or elaborate the outputs of older structures. Weights might correspond to the links or implications among processing structures.

Learning in cascade-correlation proceeds by the successive alternation of two distinct phases. The first of these, called the output training phase, consists in the adjustment of selected weights in the existing network. During this output training phase, only the weights leading to the output units are adjusted (hence the name: *output* training phase). All other weights in the network are frozen in the sense that they are not adjusted during output training. The second learning phase, called input training, is concerned with the training and installation of new hidden units into the network. During this input training phase, weights leading into units making up a pool of candidate hidden units that are separate from the network are adjusted (hence the name: *input* training phase). Only the weights leading to these candidate hidden units are adjusted during input training; all the weights in the existing network are frozen. As will be described in more detail, when a candidate unit has reached an optimal measure of performance, it is selected and installed in the existing network. Cascade-correlation then reverts back to the output training phase. The structures of some generic, hypothetical cascade-correlation networks are shown in Figure 2, for both output training (a and c) and input training (b) phases.

==========Figure 2==========

Weights are adjusted using a second-order method called quickprop (Fahlman, 1988; Fahlman & Lebiere, 1990). The quickprop algorithm is loosely based on Newton's minimization method and makes use of the current and previous derivative of the potential to be minimized in order to construct a local approximation of the potential's curvature.

then more generally, while still erring in conflict situations. Finally, they may successfully integrate the two dimensions to produce mainly correct judgments.

2.2. Previous models of balance scale phenomena

The first computational simulation of balance scale phenomena consisted of Klahr and Siegler's (1978) modeling of each the four stages in terms of production rules. This work described central features of the child's performance at each stage, but did not explain transitions between stages. Since their model appeared well before the torque difference effect was known, it did not deal with that effect. It is interesting to note that the torque difference effect is not explainable by these sorts of rules since any such rule would apply regardless of the torque difference involved in a particular problem. For example, the weight or distance on one side is greater than that on the other side regardless of how much greater it is.

The first balance scale model to address the transition issue in a serious way was by Langley (1987). He used a production system that modified its existing, overly-general rules through discrimination learning. The learning mechanism searched for differences between cases where correct predictions were made and cases where errors were made. Unfortunately, there was no detailed assessment of stages in this model. It was evaluated only by noting that there was an increased percentage of problems correct as training progressed. On the negative side, it was reported that the model learned rules that children never showed, failed to focus on weight information before distance information, and never reached stage 4. The fact that it did not focus on weight information before distance suggests that the model did not capture stages 1 and 2. It was explained that the model could not reach stage 4 since torque could not be described in the representation language that was employed and the program could not construct new representations. The model did not try for the torque difference effect and presumably could not capture it because of its exclusive reliance on symbolic rules (see General Discussion).

A rule learning program commonly used in contemporary cognitive modeling is Soar (Newell, 1990), which constructs its own rules by caching the results of look-ahead search. It too has been applied to balance scale phenomena with some success (Newell, 1990). Soar reportedly acquired stages 1, 2, and 3 but, like Langley's model, did not manage to reach the performance characteristic of rule 4. Moreover, it is unclear how dependent the Soar model was on getting balance scale problems in a certain order. It may well be that different problem orders would yield different orders of acquisition of rules. Like the other rule-based models, Soar did not try for the torque difference effect, nor is it apparent how it could in principle capture this effect.

McClelland (1989) reported a simulation of balance scale stages using a connectionist network with the back-propagation learning rule. His model required a number of limiting assumptions, including a strong bias in the training patterns favoring equal distance problems (i.e., *balance* and *weight* problems) and a forced segregation of weight vs. distance information in connections to the hidden units. The network did progress through the first three stages of the balance scale, but there was a great deal of shifting back and forth between rules 3 and 4, with stage 4 never being clearly established. Our own experimentation with McClelland's network indicates that it can reach stage 4 only by sacrificing stages 1 and 2, and that it can simulate the torque difference effect (Schmidt & Shultz, 1991).

In this paper, we report on four simulations with balance scale phenomena. The first three involve cascade-correlation networks: one focuses on stage progressions, another on the impact of different diagnostic criteria on stage assessment, and the third on the torque difference effect. The fourth simulation examines the ability of some simple symbolic rules (not Siegler's initial four rules) to capture the torque difference effect.

2. The balance scale

An emerging benchmark for detailed computational modeling in cognitive development is the ability to capture psychological phenomena associated with the balance scale. The clarity and replicability of balance scale phenomena with children, coupled with the classical developmental appeal of its stage-like character, have led to both rule-based (Klahr & Siegler, 1978; Langley, 1987; Newell, 1990) and connectionist (McClelland, 1989) models.

2.1. Psychology of the balance scale

Psychological researchers typically present the child with a rigid balance beam in which differing numbers of equal weights are placed on pegs at various distances to the left or right of a fulcrum. The child's task is to predict which side of the scale will drop when supporting blocks are removed. A five position, five weight version of the balance scale is presented in Figure 1. Typically, all of the weights on one side of the fulcrum are placed on a single peg.

==========Figure 1==========

Siegler (1976, 1981) has used the six different types of balance scale problems shown in Figure 1 to assess the rules that children might be using on this task. So-called *balance* problems have equal numbers of weights placed at equal distances from the fulcrum so that the scale balances. For *weight* problems, the side with more weights goes down since the distances from the fulcrum are equal. In *distance* problems, the side with the weights placed a greater distance from the fulcrum goes down since the two sides have equal weights. The three types of *conflict* problems have more weight on one side but more distance on the other side. The side that actually goes down is the side with greater weight for *conflict-weight* problems, and the side with greater distance for *conflict-distance* problems. The scale balances in *conflict-balance* problems.

Siegler (1976, 1981) has found that children's performance on the balance scale progresses through four distinct stages, each of which can be characterized by a symbolic rule: (1) use weight information alone to determine if the scale will balance, (2) emphasize weight information, but also use distance information in the event that the weights to the left and right of the fulcrum are equal, (3) consider both weight and distance information for simple problems, but get confused when weight conflicts with distance, (4) multiply distance by weight for each side and compare the products. Siegler has noted that each of these rules makes specific predictions about the kinds of problems that children using the rule will solve. These predictions are given by the predicted percentages correct in Figure 1. This orderly stage progression constitutes the first major psychological regularity in the balance scale literature.

The other major balance scale phenomenon is the *torque difference effect* (Ferretti & Butterfield, 1986). The torque on each side of the fulcrum is defined as the product of weight and distance for that side. The torque difference for the problem is the absolute difference between the torques on the two sides of the fulcrum. The psychological result is that the larger the torque difference, the easier the problem is for children to solve. This could be regarded as an effect of information salience; the more perceptually salient the critical information, the easier the problem is to solve.

What makes the balance scale task even more interesting is that it is an instance of a much larger class of problems in which children integrate information across two dimensions. The literature on this class of problems includes projection of shadows, inclined planes, conservation, class inclusion, fullness, and several other problems (Siegler, 1991). In each case, younger children appear to base their judgments on a single dimension, initially ignoring the other relevant dimension. Then they gradually start using the second dimension, first in restricted situations and

Machine Learning, 16, 57-86 (1994)

Modeling Cognitive Development on Balance Scale Phenomena

THOMAS R. SHULTZ, DENIS MARESCHAL,* AND WILLIAM C. SCHMIDT

Department of Psychology and McGill Cognitive Science Centre, McGill University

Abstract. We used cascade-correlation to model human cognitive development on a well studied psychological task, the balance scale. In balance scale experiments, the child is asked to predict the outcome of placing certain numbers of equal weights at various distances to the left or right of a fulcrum. Both stage progressions and information salience effects have been found with children on this task. Cascade-correlation is a generative connectionist algorithm that constructs its own network topology as it learns. Cascade-correlation networks provided better fits to these human data than did previous models, whether rule-based or connectionist. The network model was used to generate a variety of novel predictions for psychological research.

Keywords: cognitive development, balance scale, connectionist learning, cascade-correlation

1. Introduction

Although connectionist network models have become well known for their ability to simulate low level perceptual, learning, and memory phenomena, it has been unclear whether they would be suitable for modeling aspects of higher level psychological processes and their development. The recent appearance of a variety of interesting connectionist models of human development suggests some degree of applicability (Chauvin 1989; Elman, 1991; Harnad, Hanson, & Lubin, 1991; MacWhinney, Leinbach, Taraban, & McDonald, 1989; McClelland, 1989; Plunkett & Marchman, 1991; Schyns, 1991).

In addition to these new empirical results with connectionist modeling, a number of recent theoretical papers have argued that the application of connectionist models to cognitive development has fostered a return to the long neglected, but traditional concerns of developmental transition (Bates & Elman, 1993; Plunkett & Sinha, 1992; Shultz, 1991). The twin issues of structure and transition have tended to dominate developmental psychology. Whereas structural issues concern the description and diagnosis of abilities at various stages, transition issues concern the mechanisms by which the child moves from one stage to the next. Because transition has proven to be such a difficult problem, developmental psychologists have tended to ignore it in favor of more tractable diagnostic studies of children's cognition. Likewise, cognitive modelers have typically had greater success modeling processing at various stages than with transitions between stages.

In the present paper, we report on a connectionist model of cognitive development on balance scale phenomena, emphasizing both structural and transition issues.

---

*DENIS MARESCHAL IS NOW AT THE DEPARTMENT OF EXPERIMENTAL PSYCHOLOGY, UNIVERSITY OF OXFORD. WILLIAM SCHMIDT IS NOW AT THE DEPARTMENT OF PSYCHOLOGY, UNIVERSITY OF WESTERN ONTARIO.