Decreasing the number of training presentations in all models caused slower convergence, as did increasing training bias. The precise effects of segregating hidden units depended on the method of weight updating, but the general principle was that psychological realism followed slow convergence.

One ramification of the current findings is that models, like humans, need not have access to all of the information about a problem in order to succeed in finding a solution. Indeed, if models are supplied with complete information, realistic effects do not occur.

Shultz (1991) suggested that stages would emerge whenever network models solve part of the overall problem before solving the range of possible problem types. Among the techniques he listed for encouraging partial solutions (and thus stages) were hidden unit herding, over-generalization, training pattern bias, and hidden unit recruitment. Working on too much of the problem at once may encourage overly rapid convergence on a general solution and thus preclude the appearance of stages. The present findings would suggest that all of these methods slow network convergence. A useful heuristic to apply in the creation of connectionist models of cognitive development may be to consider possible convergence slowing assumptions that bear on the problem domain.

A phenomenon that may be related to the current findings is Elman's (1991) "starting small" effect. Elman reported that recurrent networks had difficulty learning a small grammar unless there was a gradual increase in either the complexity of the training instances or the "working memory capacity" of the network. Here we find analogously that models trained with a reduced number of exemplars perform more realistically, and in the case of batch updating under a heavy bias, often fail to discover a solution unless trained with a small subset of training examples. An important issue for future investigation is whether the staging of the child's environment and her developing cognitive resources work in concert to selectively filter information accessible to learning.

A second result of the current work is that McClelland's (1988) specific set of assumptions is one of several sufficiently capable of producing realistic psychological performance. Although the incremental nature of connectionist learning is crucial for the success of balance scale models (Schmidt, 1991), so is convergence slowing.

It would appear that some assumptions of McClelland's original model are replaceable. The current findings demonstrate balance scale performance without the architectural assumption of a segregated network topology. The bias and subset training assumptions, too, can be replaced by other assumptions which favor one problem dimension over another. Using a generative algorithm, Schmidt (1991) demonstrated that the state of the initial weights can place networks in a position from which they traverse the psychological rules in a realistic fashion. In another simulation, a deliberate patterning of noise added to the training set also achieved the same end.

Another generative connectionist model of balance scale phenomena also demonstrated the disposability of the segregated architectural assumption (Shultz & Schmidt, 1991). In addition, that model showed that a randomly changing environment of training instances could be replaced with a more stable, gradually expanding set of exemplars. An important issue for future research is to examine the plausibility of various sets of balance scale model assumptions and the model's corresponding ability to fit human data.

Acknowledgement

This research was supported by a grant from the Natural Sciences and Engineering Research Council of Canada.

References

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

McClelland, J. L. 1988. Parallel distributed processing: Implications for cognition and development. Technical Report AIP-47, Department of Psychology, Carnegie-Mellon University, Pittsburgh, PA.

Schmidt, W. C. 1991. Connectionist models of balance scale phenomena. Unpublished Honours Thesis, Department of Psychology, McGill University, Montréal, Québec, Canada.

Shultz, T. R. 1991. Simulating stages of human cognitive development with connectionist models. In L. Birnbaum and G. Collins eds., *Machine learning: Proceedings of the Eighth International Workshop*, pp. 105-109. San Mateo, CA: Morgan Kaufman.

Shultz, T. R. and Schmidt, W. C. 1991. A Cascade-Correlation model of balance scale phenomena. In: *Proceedings of the Thirteenth Annual Conference of the Cognitive Science Society, pp. 635-640*. Hillsdale, NJ: 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.

segregated network, continuous updating produced more runs which fit the psychological data. Figure 6 plots this interaction. Interestingly, those runs presented with fewer training instances, large biases, and trained with continuous weight updates yielded the greatest proportion of psychologically realistic results, even outperforming McClelland's (1988) original model.

Thus, the second simulation demonstrated that the earlier failures of networks to converge were due to the use of batch weight updating. The first two simulations suggest that a biased training environment and small subset training method slow convergence and enhance psychological realism.

Simulation 3: Fully Connected Nets

The final simulation investigated the effect that segregating the weight and distance dimensions had on producing psychologically realistic performance. This simulation repeated the manipulations of the first two, but without the assumption of segregated hidden units. The network topology had 20 inputs, 4 hidden units, 2 outputs, and was fully connected. Networks were trained under conditions of both continuous and batch weight updating. Since in the previous simulations a lack of training set bias did not result in rule use, this group was dropped. All other details remained unchanged from the earlier simulations.

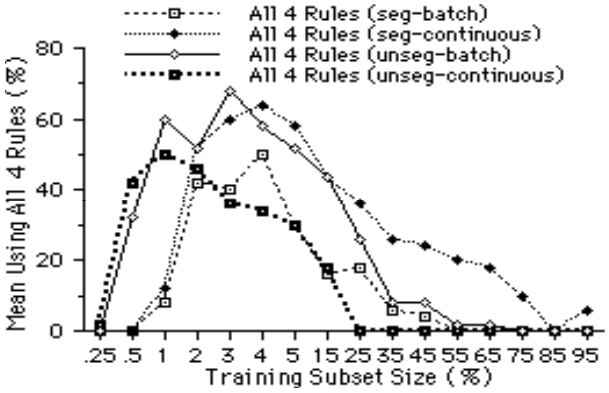

Figure 6: Psychological performance for different network topologies and weight update methods.

For each of the weight update methods (batch and continuous), segregated data generated in earlier simulations were contrasted with the new non-segregated data in 2 x 16 (network topology x subset size) ANOVAs for convergence rate and all 4 psychological rules.

For continuous weight updating, main effects were found for topology ($F(1,1248) = 54.8$, $p<.0001$), subset size, ($F(15,1248) = 866.8$, $p<.0001$) and their interaction ($F(15,1248) = 85.3$, $p<.0001$). Investigation of the interaction revealed that segregated networks converged more slowly than non-segregated for subset levels above 1%. The opposite occurred below 1%. Main effects were observed on all 4 rules for topology ($F(1,1248) = 44.9$, $p<.0001$), subset size ($F(15,1248) = 31.4$, $p<.0001$), and their interaction ($F(15,1248) = 12.5$, $p<.0001$). Investigation of the interaction showed that the segregated networks outperformed the non-segregated networks at subset sizes above 1%. Below 1%, the opposite occurred. Performance corresponded with rate of convergence, in that the interaction effects mirror one another. The group of runs with slowest convergence were also those with highest psychological performance (see Figure 6).

The slower convergence witnessed for the segregated networks seems to be a result of using fewer weights to encode the same amount of information as in the non-segregated networks. More weight changes per epoch in the non-segregated networks speeds convergence.

For batch weight updating, main effects were found for topology ($F(1,1248) = 80.7$, $p<.0001$), subset size, ($F(15,1248) = 190.7$, $p<.0001$) and their interaction ($F(15,1248) = 8.1$, $p<.0001$). Investigation of the interaction showed non-segregated networks converged more slowly than segregated, at all levels of subset size except from 2% to 5%. Main effects were observed on all 4 rules for topology ($F(1,1248) = 70.0$, $p<.0001$), subset size ($F(15,1248) = 49.4$, $p<.0001$), and their interaction ($F(15,1248) = 6.8$, $p<.0001$). Investigation of the interactions showed that the non-segregated batch networks outperformed the segregated batch networks at all levels of subset size (see Figure 6).

Thus, the final simulation showed that, with continuous weight updating, segregated networks converged more slowly than non-segregated networks, and also displayed more realistic psychological performance. With batch weight updating, the opposite effect occurred: non-segregated networks converged more slowly than segregated networks, and also showed more realistic psychological performance. In both cases, whenever network topology impeded convergence of the back-propagation learning algorithm, more realistic psychological performance followed.

The slower convergence for non-segregated batch networks may be due to the failure of so many segregated batch networks to learn. Recall that these nets tended to converge quickly on defective solutions.

Segregated networks do not invariably improve the fit to psychological data. Rather, when segregation slows convergence, as with continuous updating, better psychological performance follows; when segregation speeds convergence, as with batch updating, nets diverge from psychological realism.

Discussion

In these simulations, psychological success of the balance scale models increased as convergence slowed.

bias revealed that the more bias used in training, the slower the networks were to converge.

In every ANOVA, the interaction effects demonstrated that networks trained with different levels of bias behaved similarly at those psychologically optimal levels of small subset size. For subsets larger than 15%, increasing bias played an increasingly prominent role in preventing convergence. For subsets smaller than 0.5%, there was a gradual drop off in the magnitude of the log coefficient and in the proportion of error reduced. Convergence failed to occur for many runs at the smallest level of subset size.

Thus, the first simulation showed that McClelland's (1988) assumptions of a strongly biased training environment and of a small subset size impeded convergence of the back-propagation learning algorithm. By analyzing a wider range of these variables than were used in his model, we discovered that the most psychologically realistic data were generated by models exhibiting a slow rate of convergence. We also found a failure of back-propagation to learn successfully when trained with a bias and large subset sizes.

Simulation 2: Continuous Weight Updating

The first simulation was surprising in that so many networks failed to converge at all. To determine whether these results might be due to the use of batch weight updating, we repeated the above simulation using a continuous[3] weight updating method. A permuted presentation of training instances was used to prevent any unforeseen side effects due to auto-correlation of the sequence of exemplars.

As before, a 5 x 16 (bias x subset sizes) ANOVA of convergence rate was undertaken, revealing main effects for bias ($F(4,720) = 15.4$, $p<.0001$), subset sizes ($F(15,720) = 1426.4$, $p<.0001$), and their interaction ($F(60,720) = 39.5$, $p<.0001$). The ANOVA of learning depth also revealed main effects for bias ($F(4,720) = 180.4$, $p<.0001$), subset sizes ($F(15,720) = 3891.3$, $p<.0001$), and their interaction ($F(60,720) = 22.1$, $p<.0001$). Finally, the ANOVA predicting psychological performance demonstrated main effects for bias ($F(4,720) = 106.2$, $p<.0001$), subset sizes ($F(15,720) = 33.6$, $p<.0001$), and their interaction ($F(60,720) = 6.9$, $p<.0001$).

The convergence rate and learning depth interaction effects were negligible, and none were of interest. The average regression captured 57% of available variance ($R^2 = .57$, $sd = .24$). After excluding models which failed to converge, this average fit was 58% of the variance ($R^2 = .58$, $sd = .24$). Only 4 of the 800 (.05%) runs failed to converge, and all of these networks were at the lowest level of subset size, having been delivered too few training exemplars.

Figure 5 plots learning depth as a function of subset size for both continuous and batch updating methods. Continuous, but not batch, updating confirmed our intuitions that the amount of error reduced was proportional to the number of observed training instances.

The failure of batch weight updating to learn in a reasonable amount of time may be related to the anecdotal report that highly redundant data sets result in slower convergence on a solution with batch, but not with continuous, weight updating (see connectionists e-mail list exchanges in October 1991).

Psychologically realistic data were generated by continuous runs trained at practically all levels of subset size. However, the interaction effect for both measures of psychological performance showed that the best performance came from models with subset sizes between 1% and 5%. No models trained without bias exhibited realistic performance. A strong linear trend ($F(1,794) = 404.0$, $p<.0001$) in the cell means of bias demonstrated that the larger the bias level used in training, the more likely one was to observe psychologically realistic runs. Additionally, a weaker, but still significant linear trend among the cell means of subset size ($F(1,783) = 32.6$, $p<.0001$) demonstrated that the smaller the subset size, the more psychologically realistic the model.

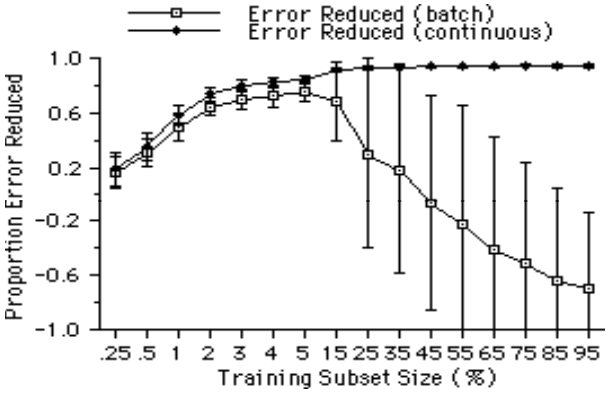

Figure 5. Learning depth for weight update methods.

Contrasting batch updating with the current data in a 16 x 2 (subset size x updating method) ANOVA on use of all 4 rules revealed main effects for subset size ($F(15,1568) = 60.7$, $p<.0001$), training method ($F(1,1568) = 9.5$, $p<.0001$), and their interaction ($F(15,1568) = 1.9$, $p<.03$). Investigating the interaction effect showed that for all levels of subset size in the

[3] Continuous (also known as per-sample, on-line, or pattern) weight updating computes derivatives and weight changes after the presentation of each pattern, as opposed to a batch (also known as per-epoch, or epoch) updating method in which the derivative of the error function summed over all patterns is taken each epoch, before weight changes occur.

A 5 x 16 (bias x subset sizes) ANOVA of the log coefficients was undertaken, revealing main effects for bias ($F(4,720) = 148.4$, $p<.0001$), subset sizes ($F(15,720) = 237.3$, $p<.0001$), and their interaction ($F(60,720) = 16.9$, $p<.0001$). A second ANOVA of learning depth revealed main effects for bias ($F(4,720) = 432.5$, $p<.0001$), subset sizes ($F(15,720) = 278.5$, $p<.0001$), and their interaction ($F(60,720) = 39.6$, $p<.0001$). A third ANOVA of the proportion of networks showing realistic psychological performance showed main effects for bias ($F(4,720) = 26.4$, $p<.0001$), subset sizes ($F(15,720) = 25.3$, $p<.0001$), and their interaction ($F(60,720) = 5.4$, $p<.0001$). In this initial simulation, the average regression captured 57% of available variance ($R^2 = .57$, $sd = .28$). After excluding models which failed to learn, this average fit increased to 73% of the variance ($R^2 = .73$, $sd = .16$).

A plot of the mean log coefficient as a function of subset size appears in Figure 3, along with the mean proportion of runs demonstrating psychologically realistic stages.[2] The shallowest learning curves occurred for networks trained with subsets of randomly chosen training instances in the range of 1% to 5%. This effect turns around below the 1% level as the learning curves begin to steepen. Could these networks trained on so few instances actually have converged faster? No, since investigation of the amount of error reduced (Figure 4) at these levels indicates that these networks failed to solve the problem at all!

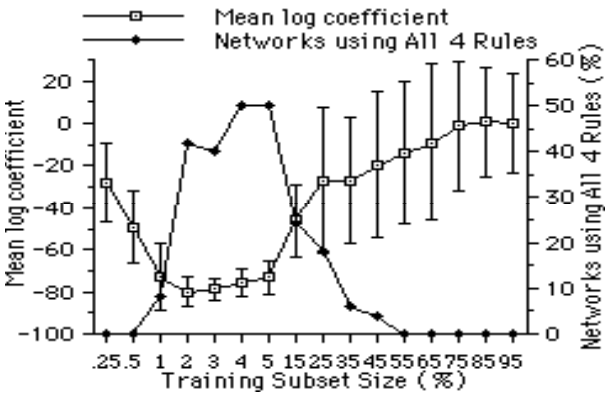

Figure 3. Mean log coefficient and percent runs showing all 4 rules.

Figure 3 also shows that the most psychologically realistic data were generated by models trained using subset sizes in the range of 2% to 25%. For both convergence rate and depth of learning, there is wider variation among networks outside this range of small subsets. This wide variation reflects the fact that fewer networks outside of this range converged on a solution. All networks which failed to converge were trained with subset sizes outside of the range of the range of 0.5% to 15%.

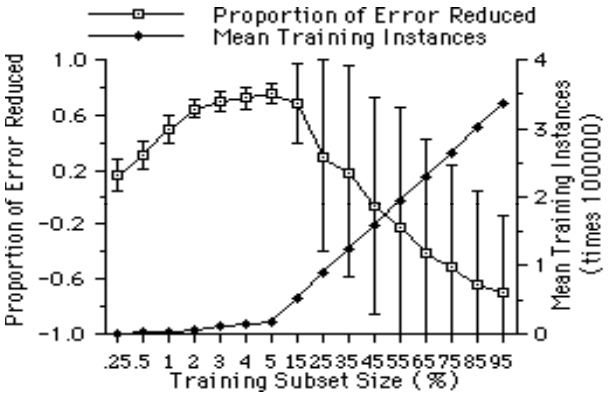

Figure 4. Mean learning depth and number of training presentations.

Investigation of the interaction effects for psychological performance revealed that models trained under levels of 5, 10, 15 and 20 times bias performed realistically while no models run without bias did. This is consistent with McClelland's (1988) two successful models, with subset size 6%, bias 5 times, and subset size 9%, bias 10 times. Here, the ordering of the cell means saw the level of 10 times bias yield the greatest proportion of psychologically realistic subjects, followed by levels of 15, 20, and 5 times bias respectively.

A plot of learning depth appears in Figure 4. Maximum error reduction occurred for networks trained with subsets of instances in the range of 2% to 15%, which overlaps with networks having the slowest convergence rates and the most psychologically realistic performances. In addition, Figure 4 plots the mean number of training instances witnessed in the models' 200 epoch lifetimes.

Comparing across Figures 3 and 4, it is clear that the number of training instances is negatively related to learning depth and positively related to the log coefficient. The negative relation of number of training instances to learning depth is an artifact of the failure of many networks to learn at high levels of subset size. The positive relation of the number of training presentations to log coefficient demonstrates that rate of convergence was generally faster for networks that had more chances to reduce error. Together this reveals that networks seeing a large number of biased training presentations converged quickly on inadequate solutions.

Unfortunately the inequality of the number of training instances across levels of bias prevented us from properly assessing the effect that bias has on convergence rate. A separate experiment controlling for the number of training presentations across all levels of

[2] All error bars in this paper represent 1 standard deviation.

interpreted to transform the network's output (2 real numbers between 0.0 and 1.0) into one of three possible prediction responses.

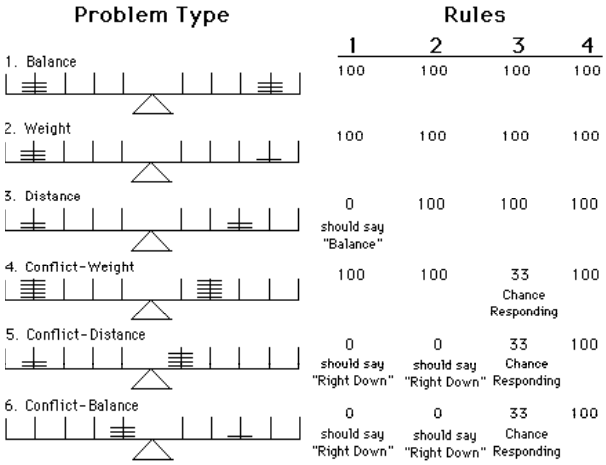

Figure 1. Predictions of percent problems correct for children using different rules.

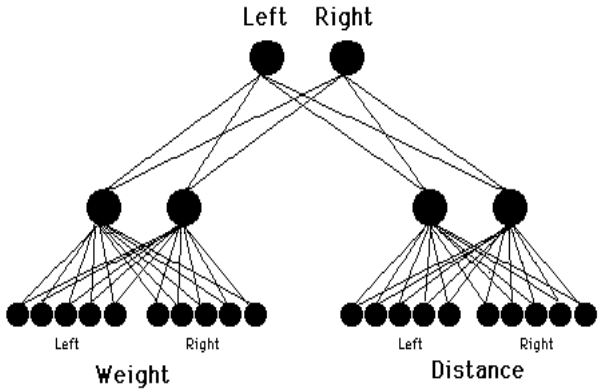

Figure 2. Network topology from McClelland (1988).

The model was trained using the back-propagation learning algorithm.[1] Learning in each epoch was from 100 instances randomly selected from the entire set of 625 possible training problems augmented with a bias for *balance* and *distance* problems (*equal-distance* problems). The bias increased the training set to include 5 times or 10 times the original number of *equal-distance* problems. After each epoch the model's performance was evaluated using Siegler's rule assessment methodology. Other than where noted, all simulations in the current paper assume the network topology, training method, and parameter settings used in McClelland's (1988) original simulations.

Simulation 1: Subset Size and Bias

The first simulation examined the effects of two manipulations. One manipulation varied the size of the bias for *equal-distance* problems. Five different levels of bias were employed. The unadulterated set of 625 possible balance scale problems was augmented with 0, 5, 10, 15 and 20 times the normal number of *equal-distance* problems, resulting in new training corpuses of 625, 1125, 1750, 2375, and 3000 patterns respectively.

The other manipulation varied the size of the subset of training instances randomly selected each epoch from the training corpus. These subsets were selected without replacement, and a permuted batch weight updating method was used as in McClelland's (1988) original model. Since the sizes of training corpuses were unequal across different levels of bias, a percentage of the total number of exemplars belonging to each training corpus was selected. The levels of subset size investigated were 0.25%, 0.5%, 1%, 2%, 3%, 4%, 5%, 15%, 25%, 35%, 45%, 55%, 65%, 75%, 85%, and 95% of the entire training corpus.

Ten runs for each of the 80 groups (5 levels of bias x 16 subset sizes) were carried out. Each run was tested on the entire set of 625 possible training patterns both before training began and after weight updating. The patterns' total sum of squared errors score (TSS) was recorded every epoch, and the network's responses were evaluated for their fit to any of the 4 psychological rules. This longitudinal rule record was assessed to determine whether or not the network passed through all 4 stages. Training continued for 200 epochs each run.

In order to evaluate the style of learning characteristic to each group of runs, a simple linear regression model was fitted to the longitudinal TSS error scores for each run, with epoch predicting error score. This yielded regression equations of the form $error = b_0 + b_1 * log(epoch)$. The log coefficient b_1 assesses the fall off of the learning curve, and hence the rate of convergence. This measure of convergence rate will be more negative for networks which reduce error more slowly. The constant b_0 offsets the learning curve from the abscissa. In the case that there is both a large b_0 term, and a small value of b_1, the network will have failed to converge.

Additionally, for each run, the proportion of error reduction over the 200 training epochs was assessed by dividing the initial TSS error less the average error score for the last 10 epochs of training by the initial level of TSS error. This value can be negative in the event that TSS error increases. Proportion of error reduction was used to assess depth of learning. A network which has a steep convergence rate, but has reduced little error, has failed to solve the problem.

[1] A batch updating method was used in conjunction with permuted presentation of training instances, a learning rate of 0.075, and momentum of 0.9. Weights in the model were initialized randomly in the range of $-0.5 \leq w_i \leq 0.5$.

Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society 1992. Hillsdale, NJ: Erlbaum.

An Investigation of Balance Scale Success

William C. Schmidt and Thomas R. Shultz
Department of Psychology
McGill University
1205 Penfield Avenue
Montréal, Québec, Canada H3A 1B1
schmidt@lima.psych.mcgill.ca || shultz@psych.mcgill.ca

Abstract

The success of a connectionist model of cognitive development on the balance scale task is due to manipulations which impede convergence of the back-propagation learning algorithm. The model was trained at different levels of a biased training environment with exposure to a varied number of training instances. The effects of weight updating method and modifying the network topology were also examined. In all cases in which these manipulations caused a decrease in convergence rate, there was an increase in the proportion of psychologically realistic runs. We conclude that incremental connectionist learning is not sufficient for producing psychologically successful connectionist balance scale models, but must be accompanied by a slowing of convergence.

Introduction

Connectionist learning algorithms have successfully acted as transition mechanisms in a number of recent models of cognitive developmental phenomena. McClelland (1988) suggested that gradual, incremental error reduction is a key property of connectionism that is responsible for this success. In the current paper, we focus on McClelland's (1988) connectionist model of cognitive development on the Piagetian balance scale task. We show that variants of the original model perform well psychologically as long as they delay convergence of the back-propagation learning algorithm.

The Balance Scale Task

The balance scale task consists of showing a child a balance scale (Figure 1) supported by blocks so that it stays in the balanced position. A number of weights are placed on one of a number of evenly spaced pegs on each side of the fulcrum, and it becomes the child's task to predict which arm will go down, or whether the scale will balance, once the supporting blocks are removed.

Siegler (1976, 1981) has reported that children's performance on the balance scale progresses through 4 distinct rule based stages: (1) use only weight information to determine if the scale will balance, (2) emphasize weight information but consider distance if weights on either side of the fulcrum are equal, (3) correctly integrate both weight and distance information for simple problems, but respond indecisively when one arm has greater weight and the other greater distance, (4) correctly integrate weight and distance information.

Siegler (1976, 1981) partitioned the set of all possible balance scale problems into 6 distinct problem types. *Balance* problems have equal numbers of weights placed at equal distances from the fulcrum. In *weight* problems, distances on either side of the fulcrum are equal, hence the side with more weights goes down. In *distance* problems, the arm with greater distance goes down since the sides have equal weights. *Conflict* problems have greater weight on one arm and greater distance on the other. The correct response to the problem determines its classification as a *conflict-weight*, *conflict-distance*, or *conflict-balance* problem. Performance in terms of the percentage of correct predictions made on some subset of problems drawn from each of the problem types can be used to classify subjects as conforming to a particular rule. Rules and their predicted performance levels for each of the 6 problem types appear in Figure 1. In order to classify children's performance, Siegler (1981) used 24 testing instances (4 from each of the 6 different problem types). Children scoring 4 or fewer deviations from responses predicted by a given rule were counted as acting in accord with the rule. Additionally, Siegler introduced a number of safeguards to ensure that a child classified at one stage was not actually responding in a manner characterized by another.

McClelland (1988) reported the creation of a connectionist model of the balance scale task with 5 pegs and 5 weights per arm. The network topology, which appears in Figure 2, consisted of two sets of 10 input units, each fully connected to a distinct pair of hidden units. Each pair of hidden units was fully connected to two output units. One set of input units represented, in a localist fashion, the number of weights on each of the balance scale's two arms (1 to 5), while the other represented the distance of the weight from the fulcrum (1 to 5). Activation values of the outputs were