A Connectionist Model of a Continuous Developmental Transition in the Balance Scale Task

Anna C. Schapiro

Honors Thesis, Symbolic Systems

2009

I certify that this honors thesis is in my opinion fully adequate in scope and quantity to meet the standards for graduation with honors in Symbolic Systems.

Advisor: James McClelland, Department of Psychology, Stanford University

Signed: Date:

I certify that this honors thesis is in my opinion fully adequate in scope and quantity to meet the standards for graduation with honors in Symbolic Systems.

Second reader: Anthony Wagner, Department of Psychology, Stanford University

Signed: Date:

Table of Contents

Abstract 6
1. Introduction 7
    1.1 Balance scale task 8
2. Experiment 1 of Jansen and van der Maas (2001) 14
    2.1 Bimodality and inaccessibility 14
    2.2 Hysteresis and sudden jump 15
    2.3 Divergence 16
    2.4 Summary of results for children's behavior 17
    2.5 Latent class analysis 19
3. Simulations of Experiment 1 using the McClelland (1989) model 21
    3.1 Representation of the task 21
    3.2 Training 23
    3.3 Testing 24
    3.4 Results 25
    3.5 Discussion 27
4. Extensions to the model 29
    4.1 Adaptive modification of gain 30
    4.2 Details of gain adjustment procedure 31
    4.3 Noise 32
    4.4 Parameters and simulation details 33
    4.5 Results 34
5. General discussion 41
6. Conclusions 49
Acknowledgments 51
Appendix A. Comments on latent class analysis 52
Appendix B. Details of network testing items 56
Appendix C. Explored extensions 57
    C.1 Learning during test 57
    C.2 Forcing symmetry 58
    C.3 Forcing symmetry by symmetrizing the test set 58
    C.4 Weight slaving 59
    C.5 4-Input architecture 60
References 62

Abstract

A connectionist model of the balance scale task is presented which exhibits developmental transitions between 'Rule I' and 'Rule II' behavior (Siegler, 1976) as well as the 'catastrophe flags' seen in data from Jansen & van der Maas (2001). The model extends the McClelland (1989, 1995) model of this task by introducing intrinsic variability into processing and by allowing the network to adapt during testing in response to its own outputs. The simulations direct attention to several aspects of the experimental data indicating that children generally show gradual change in sensitivity to the distance dimension on the balance scale. While a few children show larger changes than are characteristic of the model, its ability to account for nearly all of the data using continuous processes is consistent with the view that the transition from Rule I to Rule II behavior is typically continuous rather than discrete in nature.

1. Introduction

What is the nature of the underlying knowledge representation that determines patterns of performance and developmental change in children? This question has inspired an enormous amount of empirical and theoretical work aimed at inferring mechanisms of development based on children's behavior. One window into these developmental mechanisms that has been used extensively is children's performance on the balance scale task. Interpretations of the data from this task have tapped into a greater debate in cognitive science between two perspectives. At one end of the spectrum is what we will call the rule-based approach, which holds that performance in tasks like the balance scale task is based on a small number of distinct and discrete rules that can be used to generate responses to test items. Development consists of a progression through the use of a sequence of these rules. In this view, children's behavior is not simply describable by rules but is actually caused by the use of explicit rule representations, e.g., through the retrieval of an explicit rule from long-term memory to be used in a task (Kerkman & Wright, 1988). At the other end of the spectrum is what we will call the continuous perspective, which holds that information is represented in a more graded manner that is only approximately characterizable by the kinds of rules in rule-based approaches, and transitions between stable stages of performance are not in fact so abrupt when considered carefully. Connectionist models provide a possible mechanism for this continuous change, in which knowledge is stored as the weights of connections between simple neuron-like processing units. Rule-like behavior in a task like the balance scale task emerges from small incremental changes in the weights between these processing units, which in turn lead to incremental changes in units' activations. In the connectionist

and, more generally, the continuous approach, apparent qualitative change need not reflect a discrete transition; behavior that might sometimes look like rule change is seen as arising from incremental change in what is underlyingly continuous processing.

Though connectionist models can approximate to an arbitrary degree of accuracy the rule-like behavior of a system that explicitly incorporates rules into knowledge representation, the rule-based and continuous approaches have different tendencies in their behavior that do not motivate identical empirical predictions. For example, especially in periods of transition, a continuous model predicts that there will tend to be graded sensitivity to the dimensions relevant to the transition, whereas a strict rule-based approach predicts that sensitivity to a particular dimension will either be present or absent. The question we address here is whether there is indeed the kind of graded sensitivity that would be expected from the kinds of transitions that tend to occur in continuous models. We consider an elaborated version of McClelland's (1989, 1995) connectionist model of the balance scale task and compare it in detail to aspects of the relevant experimental data. The model is used to account for the patterns of performance found in an extensive investigation of transitions in balance scale task performance by Jansen and van der Maas (2001), bringing out aspects of the empirical data that indicate continuity in transition.

1.1 Balance scale task

In the balance scale task, developed originally by Inhelder and Piaget (1958; Piaget & Inhelder, 1969), children are shown a balance scale with a varying number of weights placed on pegs on each side, at varying distances from the fulcrum (see Figure

1). While the movement of the scale is prevented, the children are asked to imagine what would happen if the scale were allowed to move freely; they indicate whether they think the left or right side of the scale would fall, or whether the two sides would be in balance.

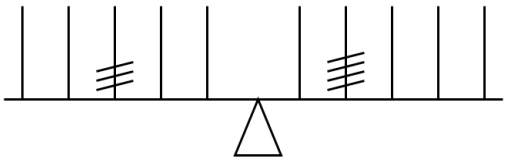


Figure 1. The type of scale used in the balance scale task.

Siegler (1976, 1981) adopted the balance-scale paradigm to test whether children's behavior on the task is best described by the use of rules. He used six item types: balance, weight, distance, conflict-weight, conflict-distance, and conflict-balance. Children were classified as using a particular rule based on their responses to test items of each of these different types. Balance items have the same number of weights at the same distance from the fulcrum on each side. Weight items have different numbers of weights on each side at the same distance from the fulcrum. Distance items have the same number of weights at different distances. Conflict items have fewer weights at a greater distance on one side of the fulcrum and more weights at a smaller distance on the other. In conflict-weight items, the correct answer is that the side with more weight falls. In conflict-distance items, the correct answer is that the side with weights at a greater distance falls. In conflict-balance items, the correct answer is that the sides balance.

Based on responses to a test containing several items of each of the above item types, Siegler (1976) claimed that children use one of four rules in determining which

side of the balance scale falls. The rules concern how to incorporate information from the two relevant dimensions of the task. The first dimension is weight, which is called the dominant dimension because younger children appear to be more sensitive to it, and the second dimension is distance from the fulcrum, identified as the subordinate dimension. According to Siegler's analysis, children using Rule I make their decision based only on the number of weights on each side of the fulcrum. Children using Rule II take distance into account when the number of weights on both sides of the fulcrum is equal; otherwise they make their decision based only on weight. Children using Rule III always consider distance and weight but "muddle through", guess, or use some other incorrect rule when the dimensions conflict. Rule IV is seen only in a minority of adolescents and adults (Siegler & Chen, 2002). Individuals using Rule IV correctly make their decision by comparing the torques (number of weights multiplied by distance of those weights from the fulcrum) on each side of the scale.

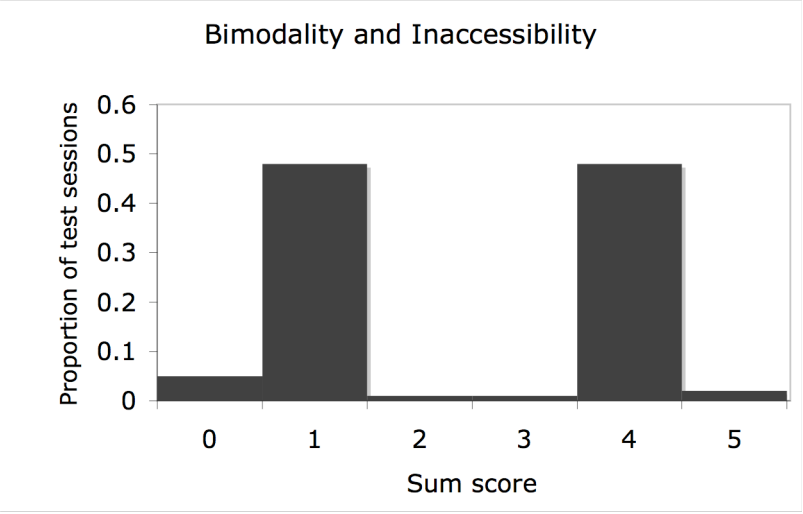
Since Siegler's seminal investigations, there have been many additional studies of the balance scale task, many of which argue for alternative characterizations of the nature of children's underlying knowledge representations. First, the existence of rules in addition to the ones originally studied by Siegler has been suggested (Boom, Hoijtink, & Kunnen, 2001; Ferretti, Butterfield, Cahn, & Kerkman, 1985; Normandeau, Larivee, Roulin, & Longeot, 1989; Jansen & van der Maas, 2002; Siegler & Chen, 1998; Van Maanen, Been, & Sitjsma, 1989). Most relevant here, though, are several instances of patterns observed in children's responses that are not fully consistent with any single rule (Jansen & van der Maas, 1997, 2002; Siegler, 1981; van der Maas & Jansen, 2003). Jansen and van der Maas explained these inconsistencies in terms of rule switching that

occurs during transition (1997) but admit that their presence is not ideal for a rule-based perspective (2002, p. 384). Another phenomenon that has been taken as evidence against the rule-based perspective is the torque difference effect (Ferretti & Butterfield, 1986, 1992; Ferretti et al., 1985). Children are more likely to behave in accordance with a more advanced rule when the difference between the torques on the two sides of the scale is greater. This result is suggestive of a continuous rather than a discrete or categorical rule-based mechanism. It has been argued that this effect only exists at extreme torque difference levels (Jansen & van der Maas, 1997; van der Maas, Quinlan, & Jansen, 2007; van Rijn, van Someren, & van der Maas, 2003), but we find the clear trend for increasing accuracy with increasing torque difference across all levels (see Figure 4, Ferretti & Butterfield, 1986) to suggest the effect is present even at low torque difference levels. Indeed, using variants of the balance scale task that allow for continuity in children's responses, Wilkening and Anderson (1982, 1991) have found direct evidence that children integrate information about weight and distance in a way that is better described by the weighted combination of the continuous dimensions of the task than by a set of discrete decision-tree rules like Siegler's.

In a series of relevant articles, Jansen and van der Maas have argued that a rule-based perspective is the best characterization of children's development on the balance scale task (1997, 2001, 2002; van der Maas & Jansen, 2003; Quinlan, van der Maas, Jansen, Booij, & Rendell, 2007). They subscribe to a definition of *rule* that requires behavior to be regular, consistent, and discontinuous (among other things; Quinlan et al., 2007). Although they generally favor a rule-based approach, they conclude that only the transition between Rule I and Rule II and the transition to Rule IV actually satisfy their

conditions (2002). One body of data often cited for the discontinuity between Rule I and Rule II is from Jansen and van der Maas (2001). The simulations in this paper will address these data specifically in order to show that the evidence is consistent with a continuous account even in this transition.

Jansen and van der Maas (2001) apply the so-called *cusp model* to data on the transition between Rule I and Rule II to test for signs of the discontinuity. The cusp model is derived from catastrophe theory, a mathematical theory intended to allow measurement of qualitative transitions, which are defined as sudden changes in a dependent variable resulting from small continuous changes in independent variables (Raijmakers, van Koten, & Molenaar, 1996). According to van der Maas and Molenaar (1992), the cusp model does not specify a mechanism for qualitative change; it only describes that change. In this application of catastrophe theory to the balance scale task, the dependent variable is interpreted as the number of correct responses to a set of distance items and the independent variables are the ability to encode the distance difference—the difference in the distance of the weights to the fulcrum on the two sides—and the number of weights placed on the balance scale on distance items. Derived from catastrophe theory are indications, called catastrophe flags, that a system is undergoing a qualitative transition. These catastrophe flags can be detected based on the behavior of the variables described above. Jansen and van der Maas (2001) investigate the presence of some of these catastrophe flags—to be described in detail below—to find evidence in children's behavior for the discontinuity in transition from Rule I to Rule II.



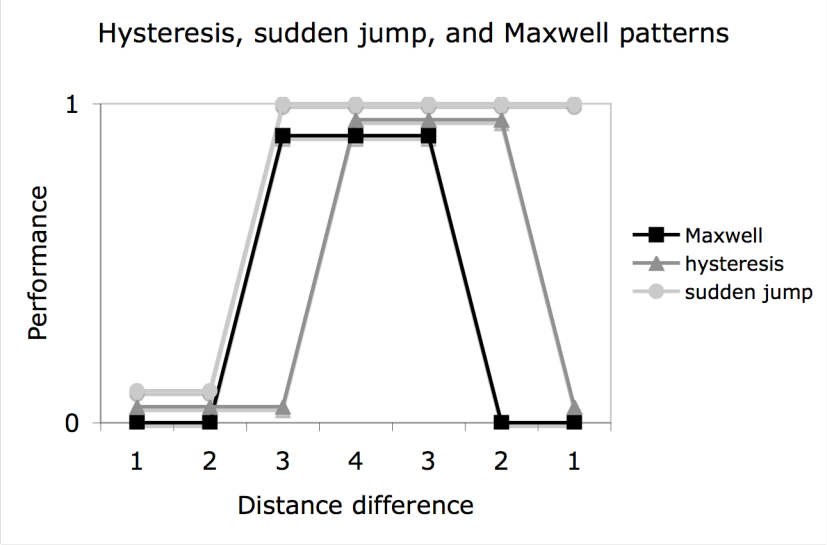


Figure 2. Examples of catastrophe flags. In the bottom graph, correct performance on the item with a given distance difference is represented as a 1 and incorrect performance as a 0. Adapted, with permission, from Fig. 2, p. 455, of Jansen and van der Maas (2001).

2. Experiment 1 of Jansen and van der Maas (2001)

In Experiment 1, Jansen and van der Maas (2001) tested 314 children from 6 to 10 years old on a paper-and-pencil version of the balance scale task using pictures of balance scales with different combinations of weights and distances (similar to Figure 1). For each item, the children had to circle an image corresponding to what they thought the scale would look like if the blocking pin preventing the scale from moving were taken out. Each child saw a total of 40 items, which were arranged into a practice test, pretest, hysteresis test, posttest, control test, and divergence test. These items were designed to test specifically for the presence of the bimodality, inaccessible region, hysteresis, sudden jump, and divergence catastrophe flags.

2.1 Bimodality and Inaccessibility

Bimodality and inaccessibility were assessed in the pretest, posttest, and divergence test. Bimodality in this context refers to a bimodal distribution of scores when some items in a test set require behavior at the level of Rule II for correct performance, and some only require behavior at the level of Rule I. In the pretest and posttest, there were three distance items, one weight item, one conflict-weight item, and one conflict-distance item. All six items in the divergence test were distance items. In the pretest and posttest, based on Siegler's (1976) rules, children using Rule I would succeed on the weight and conflict-weight items, and children using Rule II would succeed on those as well as the distance items. Conflict-weight items were not used in the analyses because scores on those items negatively correlated with the other item types. The expected distribution for scores on the pretest and posttest therefore becomes a bimodal one with

modes falling on one item correct for Rule I behavior and four items correct for Rule II behavior (see Figure 2). For the divergence test, the expected distribution has modes at zero correct and six correct, since it consists entirely of distance items. Inaccessibility refers to absence of scores in the region between these modes, which is expected in a rule-based perspective because scores between modes are inconsistent with both Rule I and Rule II behavior.

2.2 Hysteresis and Sudden Jump

Hysteresis and sudden jump were assessed in the hysteresis test. The hysteresis test consisted of a series of distance items, where the distance difference was incrementally increased and then decreased over nine items. Increasing the salience of the distance dimension was expected to cause some children using Rule I who are on the verge of transition to switch to Rule II, though most children are expected to be consistent Rule I or Rule II users. A sudden jump is characterized by an immediate shift to using Rule II as the distance difference increases, with no shift back to Rule I as the distance difference decreases again (see Figure 2). In this application of the cusp model, the child is thought to suddenly realize that the distance dimension should be considered at some point during the series of increasing distance differences because with each step the distance dimension becomes more salient and therefore easier to encode. Such a transition to the use of Rule II could then lead the child to continue to perform correctly on distance items for the rest of the hysteresis test.

Another pattern, hysteresis (also called the *delay pattern*), occurs if the child shifts to using Rule II as in the sudden jump, as the distance difference increases, and

then shifts back to using Rule I at a lower distance difference than the one at which she shifted to Rule II. The child persists in Rule II behavior until the distance dimension becomes less salient than it was when she made the switch to Rule II. In the cusp model, the presence of this pattern is considered sufficient evidence to conclude that the transition in question in discontinuous.

A third pattern a child can follow in the hysteresis test is the Maxwell convention, which is like hysteresis except that the child switches back to Rule I use at the same distance difference that she switched to Rule II use. The Maxwell convention is not considered a catastrophe flag. Patterns of this type would be expected if the child simply had a graded sensitivity to distance, allowing correct performance on items with large distance differences to coexist with incorrect 'balance' responses for small distance differences.

The control test was designed to control for the possibility that children would change their responses in the hysteresis test because they were seeing so many of the same type of item in a row. The greyness of the balance scale item weights was gradually changed from black to white and back to black over the nine items. Identical distance items were used in this test with a distance difference of three.

2.3 Divergence

The last catastrophe flag, divergence, was assessed in the divergence test. The divergence test had six items, which were all distance items with a distance difference of two. Three of the items had one weight on each side, and three had five weights on each side. The divergence hypothesis, as stated by Jansen and van der Maas, is that the

distribution of scores for the items with five weights is expected to be more bimodal than the distribution of scores for the items with one weight. They expect that children will be more likely to behave consistently with whichever rule they are using when the dominant dimension, weight, is more salient. The divergence hypothesis follows from this application of the cusp model to the balance scale task, but it does not represent the most intuitively clear sign of qualitative transition and was in fact not detected in the children's data or in any of the model simulations.

2.4 Summary of Results for Children's Behavior

The pretest, posttest, and divergence test distributions all showed the bimodality catastrophe flag as expected (see Figure 3). The modes for pretest and posttest were at scores of one and four and the modes for the divergence test were at scores of zero and six. There was also some degree of inaccessibility between the two modes, with fewer occurrences of scores that indicate behavior between Rule I and Rule II, especially in the divergence test. There was an effect of learning from the pretest to the posttest, where a significant number of children moved from scores of one on the pretest to higher scores on the posttest.

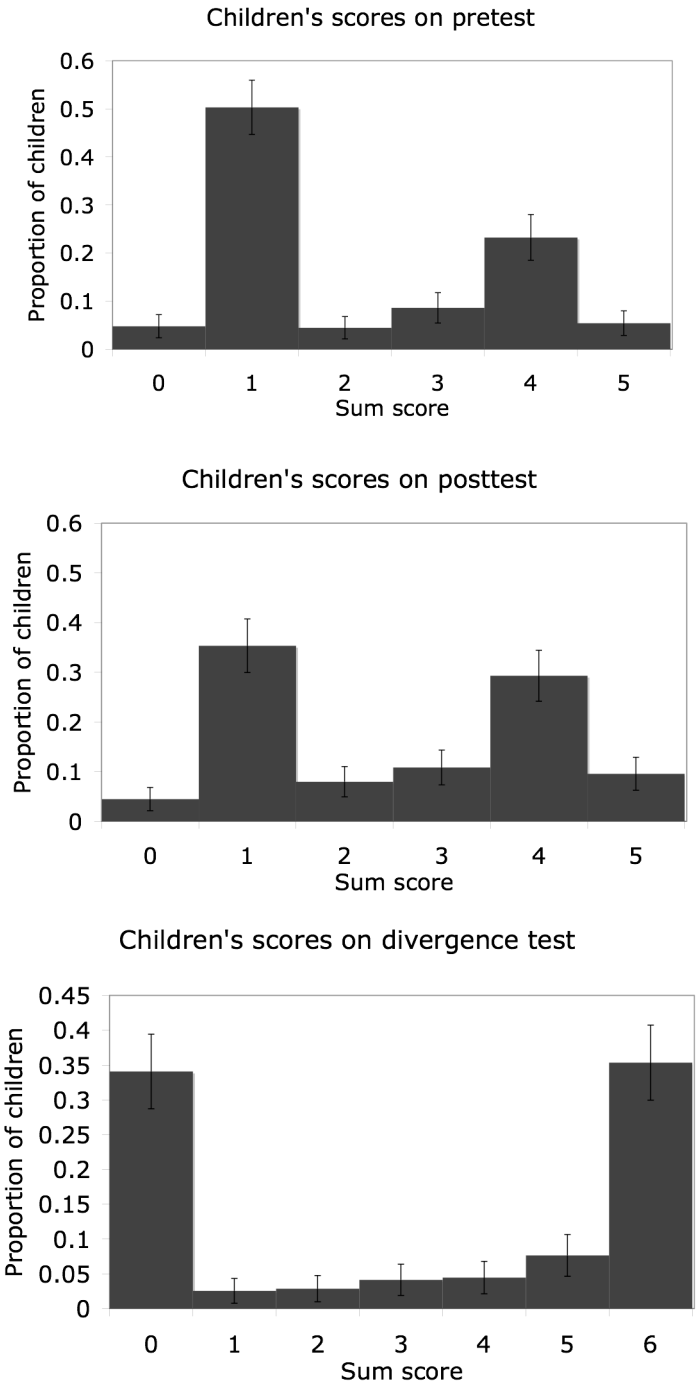


Figure 3. Results from Jansen and van der Maas (2001, Experiment 1). Distributions of children's scores on the pretest, posttest, and divergence test. A child's score for a given test was not counted in the distribution if any of the values for that test were missing.

Redrawn with permission from Figs. 5 and 6 (both on p. 475) of Jansen and van der Maas (2001).

The hysteresis and sudden jump catastrophe flags were present in the hysteresis test. Out of 314 children, 7 (2.228%) displayed a hysteresis pattern, 26 (8.280%) displayed a sudden jump pattern, and 11 (3.503%) displayed a Maxwell convention pattern. In the control item set, no children displayed hysteresis or Maxwell patterns, and 9 (2.866%) children displayed a sudden jump pattern. The divergence catastrophe flag was not found in the divergence test. The distribution of the scores for items with five weights and the distribution of scores for items with one weight were not significantly different.

In summary, four of the five catastrophe flags studied were found in the experiment: bimodality, inaccessibility, hysteresis, and sudden jump. There was also an overall learning effect from the pretest to the posttest.

2.5 Latent class analysis

In their 2001 article and other articles, Jansen and van der Maas have used latent class analysis as one of several ways of analyzing children's performance. In discussion of their findings from these studies, we do not consider the LCA results, relying instead on a direct examination of the distribution across children of specific performance profiles. We do not rely on the LCA analysis results for the following reasons: First, LCA treats the data as coming from a finite set of discrete classes, and we question the assumption that this treatment is correct. Second, in practice, the application of LCA

requires the researcher to impose constraints to limit the number of free parameters. Decisions must then be made about exactly what constraints should be imposed, and different decisions can lead to different results. The method's sensitivity to such decisions makes it possible for the method to obscure rather than reveal the structure present in the experimental data. In support of these points, Appendix A examines Jansen and van der Maas' (2001) application of LCA to their pretest and posttest data. We provide evidence that the pattern of results produced by this analysis can change if different choices are made in imposing constraints on the parameters. Our point is not to argue that LCA should not be used, but only to argue that the method is not free of difficulties (for further discussion of the strengths and weaknesses of LCA, see Shultz & Takane, 2007; Siegler & Chen, 2002; and van der Maas, Quinlan, & Jansen, 2007). Because of the difficulties with LCA, we have relied instead on a more direct consideration of the distribution of performance profiles such as those in Table 1 and Table 2.

3. Simulations of Experiment 1 using the McClelland (1989) Model

We next present a simulation of the Jansen and van der Maas experiment (2001, Experiment 1) using McClelland's (1989) model, looking specifically for the presence of bimodality, inaccessibility, hysteresis, sudden jump, and overall learning from pretest to posttest. The model has been criticized for not exhibiting the catastrophe flags seen in children's responses (Raijmakers et al., 1996), and our simulations confirm that the model falls short of accounting for many of these trends in the data. This exploration of the original model's behavior on the Jansen and van der Maas experiment will serve as a basis for understanding what behavior the original model was already capable of successfully explaining and why the extensions we later add allow a significantly better fit to the data.

3.1 Representation of the task

The network makes the same decision as the human subjects; namely, when given a scale with a certain number of weights at certain distances from the fulcrum, the network decides if the left side of the scale goes down, if the right side of the scale goes down, or if the sides balance. The network's simple three-layer architecture is shown in Figure 4.

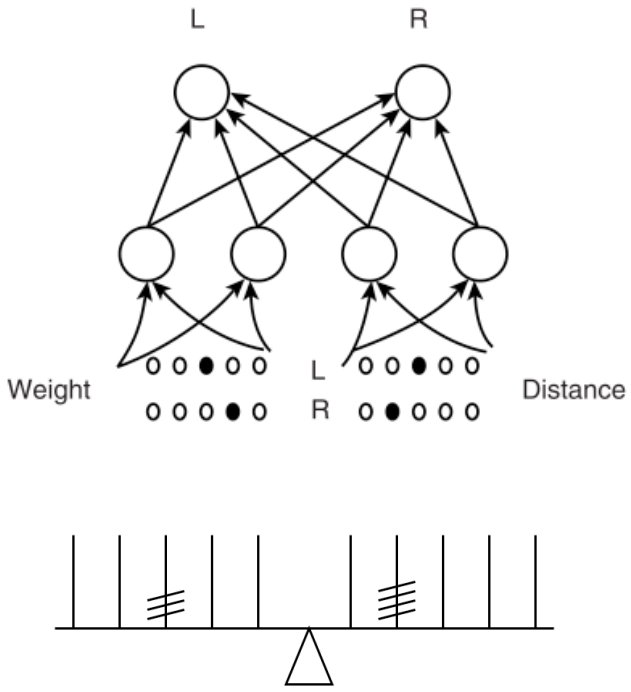


Figure 4. Network used in the McClelland (1989, 1995) model of the balance scale task. The input units that are filled in are activated to represent the network being presented with the balance scale item shown. The separation of the input units into the left and right sides of the fulcrum and the ordering of the weight and distance units from lowest to highest, as depicted, are unknown to the model before training. Reprinted from Figure 2.7 of McClelland (1989).

There are 20 input units, which are used to represent the numbers of weights and distances of those weights from the fulcrum for a given item. Each of these 20 units corresponds to a different possible weight or distance value on the right or left side of the scale. Five of the input units are used to represent one through five weights on the left side of the scale, and five other units are used to represent one through five weights on the right side of the scale. There are also five units for the five distances of the weights

from the fulcrum on the left, and five other units for the distances of the weights from the fulcrum on the right. The network does not know before training which input units correspond to which weights and distances on the scale. Each input unit has an activation of 1 when it is being used to represent its particular weight or distance value and an activation of 0 otherwise.

Each of the 10 weight units projects to two of the hidden units, and each of the 10 distance units projects to the other two hidden units. This architecture implements the assumption that weight and distance are separately assessed before they are combined when participants reason about balance scales (see McClelland, 1989, 1995, for further discussion). Each of the four hidden units projects to each of the two output units. The output units, L and R, correspond to the scale tipping to the left or to the right. The network's representation of the left side tipping would have the L output unit near an activation of 1 and the R output unit near an activation of 0, and vice versa if the right side tips. Activations near 0.5 for both L and R indicate the network's decision that the scale balances. More specifically, if the activation of the L output unit is less than 1/3, the right side is interpreted as falling. If the activation of the L output unit is greater than 2/3, the left side is interpreted as falling. Otherwise the scale is interpreted as balancing.

3.2 Training

The training set has all the possible combinations of one through five weights at one through five distances on one peg on the left and one through five weights at one through five distances on one peg on the right, for a total of 5 x 5 x 5 x 5 = 625 items. There are also nine added copies of each of the items that has weights at the same

distance from the fulcrum on each side (1125 items added, 1750 total). These copies predispose the network to treat weight as the dominant dimension. (Whether greater exposure to cases in which weight varies is in fact the true basis of the dominance of weight is not clear. Although McClelland, 1989, argued that this is one possible basis for the effect, another is that distance is a more complex relationship, depending jointly on the position of the weights and the position of the fulcrum. See McClelland, 1995, for discussion.) The weights connecting the input and hidden layers and the hidden and output layers are initialized with random values uniformly distributed between -0.5 and +0.5. The network is trained in each epoch on the entire set described above in randomly permuted order. Weights are updated after each item is presented using back-propagation. No momentum is used, and the learning rate is 0.02.

3.3 Testing

After every epoch of training, the network was tested on the items used by Jansen and van der Maas (2001), excluding items with values of six for weight or distance (see details of the items in Appendix B). These items were always presented in the same order. All of the items were identical in the control test, since the greyness of the weights is not represented within the structure of the model. As in earlier simulations with this model, the connection weights were frozen during test sessions so that there was no change in the network's response as a result of experience with test items.

The network was run independently 10 times and epochs 5 to 60 of each run were used in analysis. The test session at the end of each epoch is meant to represent an individual child doing the experiment. The range of epochs was chosen to obtain an

approximate match to the overall range of performance across the children tested in Jansen and van der Maas (2001).

3.4 Results

The data for the pretest, posttest, and divergence test items all showed bimodality. As in the data from Jansen & van der Maas (2001), the modes for the pretest and posttest distributions were at scores of one and four, and the modes for the divergence test distribution were at scores of zero and six (see Figure 5). Though there is a clear inaccessible region in the data from the divergence test, there is a less pronounced inaccessible region in the pretest and posttest. There is also no learning effect from pretest to posttest, which is expected because there is no basis for any change in performance during the test phase of the network. Accordingly, the pretest and posttest distributions are almost identical.

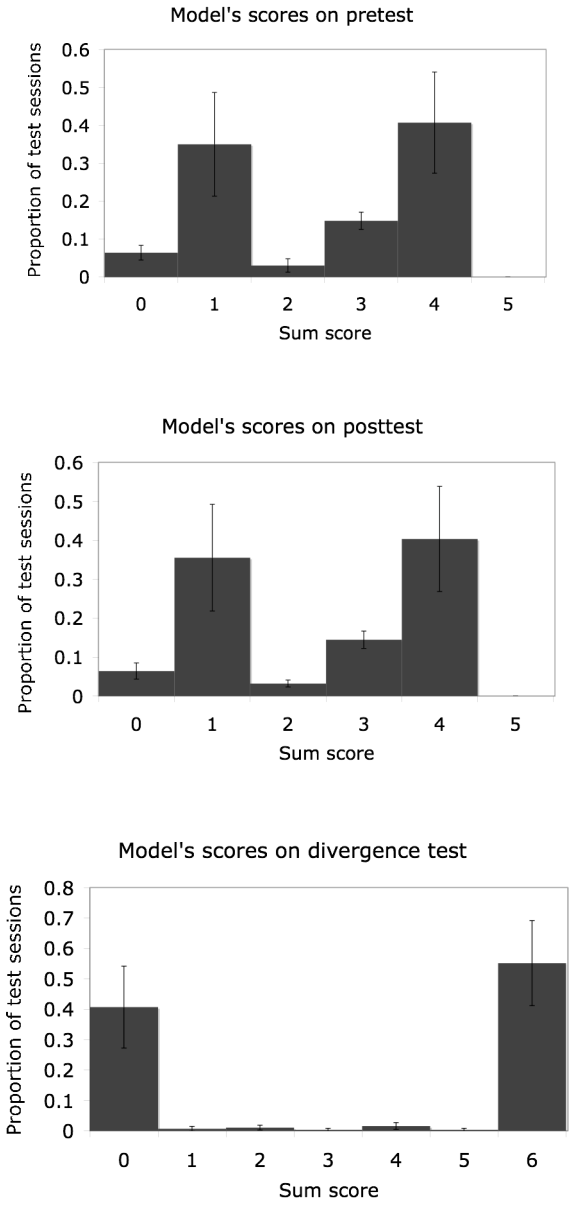


Figure 5. Results from a simulation of McClelland's (1989) model on the items from Jansen and van der Mass (2001). Distributions of pretest, posttest, and divergence test session scores. Ten independent networks were each tested after epochs 5 through 60, thereby contributing 56 scores per network. Statistical tests and confidence intervals treat the network as the random effect factor, although the random sequence of training experiences within each epoch induces considerable (though not complete) independence between epochs within networks.

The sudden jump and hysteresis catastrophe flags were not present, which is also expected because there is no basis for change in the network's performance during test. Out of 560 trials, there were no hysteresis or sudden jump patterns and 91 (16.250%) Maxwell convention patterns. In the control items, there were no hysteresis, sudden jump, or Maxwell patterns.

As in Jansen and van der Maas (2001), the divergence hypothesis was not supported by the network data. In fact, it was not supported by any simulations we ran on any version of the model. The distribution for the items with one weight and the distribution for the items with five weights were very similar [χ^2 (3, N = 560) = 2.2166, p = 0.5287].

3.5 Discussion

As in earlier investigations, the McClelland (1989) model appears to capture many of the patterns seen in children's balance scale behavior. Indeed, when there were deviations from rule-like behavior observed in children (Siegler, 1981), they tended to be similar to the types of deviations observed in the model (McClelland, 1989), a result very suggestive of the need for some continuity in the mechanisms underlying this task. In the present case, these deviations are represented by the fact that, although there is indeed bimodality and inaccessibility in children's scores on the balance scale task, neither children nor the model exhibit complete inaccessibility. The tendency of a fraction of the children and a fraction of the network testing sessions to result in intermediate scores is suggestive of the presence of at least some degree of underlying continuity in children, as in the model. The presence of Maxwell patterns in the hysteresis test is another indication

of graded sensitivity to the distance dimension, since Maxwell patterns represent better performance on items with higher distance difference. Maxwell patterns are also seen in the data from Jansen and van der Maas (2001), though not as frequently as in the simulation.

There are thus many successes of this model in describing the overall appearance of general rule-like behavior as well as the observed deviations from that behavior. The model uses no explicit representations of rules, suggesting that the patterns of children's behavior on this task may not need to be explained through use of rules. The results also suggest the stronger point that a more continuous account may be required to account for many of the details of the data, since a strict rule-based account requiring consistent performance within item types would not predict *any* scores in the 'inaccessible' region, or *any* Maxwell patterns.

Despite the successes, this model does have significant shortcomings. It is not able to show any changes in scores from pretest to posttest, and does not show the sudden jump and hysteresis catastrophe flags found by Jansen and van der Maas (2001). These transitional behaviors are important for the model to account for, and the sudden jump and hysteresis catastrophe flags in particular might seem to pose a challenge for the graded, continuous mechanisms of the model. We now present extensions to the model, however, that allow it exhibit all of these effects while still in this continuous framework, suggesting that these catastrophe flags are perhaps not so indicative of discontinuous change after all.

4. Extensions to the Model

Our approach to extending the model begins with the observation that the balance scale test situation may differ in several ways from the situations in which children learn naturalistically about balance. In naturalistic situations—for which the training regime used with the model is intended as a simplified proxy—children are thought to make implicit predictions in the course of, e.g., play on a teeter-totter at a playground. In these situations, the mismatch between the observed outcome and the implicit prediction is treated in the model as underlying gradual implicit learning over developmental time. In testing situations, however, in which a child is confronted with a long series of highly similar balance scale problems one after another, we suggest that additional processes may come into play. One of these may be the allocation of attention based on a child's own overt (and hence categorical) response to a given balance-scale stimulus configuration.

This idea has elements in common with several other approaches to performance change in the balance scale task, in which progress from 'Rule 1' to 'Rule 2' is thought of as arising from a change in the use of the distance information (c.f. van Rijn et al., 2003; Siegler 1976). Our approach differs from these other approaches, however, in that the other approaches treat the use of distance information as an all-or-nothing matter, while in our approach the influence of distance on the decision is a matter of degree, modulated up or down by attention. (As an alternative to an attention-based approach, we also considered the possibility that a child's own overt response might drive connection weight adjustment. While this remains a possible approach, our explorations of this

approach did not yield results as good as those based on the allocation of attention. See Appendix C for details of these and other explorations.)

4.1 Adaptive Modification of Gain

How might the allocation of attention be adjusted in a graded or continuous fashion, based on a network's response to a given test problem? One proposal is to use the adjustment of *gain* (Krushke, 1992; Krushke & Movellan, 1991), where gain is a parameter scaling the effect of a unit's net input on its activation. Following Krushke (1992), we adopted the idea that dimensional attention, operationalized as an adjustment to a dimension-specific gain parameter, can be adjusted using a gradient-descent procedure. To implement this within our network, we gave the distance and weight hidden unit pools separate gain values that separately control the degree of attention to the distance and weight dimensions. Specifically, the activation of each of the hidden units processing weight information was given by $a = 1/(1 + \exp(-(g_w * net)))$ where g_w is the gain parameter for the weight dimension. The activation of the hidden units for distance was similarly modulated by the gain parameter for the distance dimension.

One innovation in our procedure relative to Krushke (1992) is that instead of using the difference between an externally provided teaching signal and the activations produced by the network to calculated the necessary 'error' terms to drive attention adjustment, we instead used the difference between the network's categorical overt response and the graded activation values on which that response was based. We represented the categorical overt response in the same way that we represented externally

provided outcome information. That is, an overt response of 'left side down' was represented [1, 0]; 'right side down' as [0, 1]; and 'balance' as [.5, .5]. Informally, this approach can be thought of as implementing the idea that the gain is adjusted to make the network's response to a given pattern more definite or categorical. For example, output unit activations of [.83, .21] would be scored (as in the original model, see description above) as a left-side-down response, corresponding to the categorical response pattern [1, 0]. Adjustments would then be made to both gain parameters to reduce the difference between the assigned categorical response pattern and the underlying graded activation values.

4.2 Details of gain adjustment procedure

Gain was initialized at 1.0 at the beginning of each test session and subsequently adjusted using the difference between the network's actual graded response and the discretized response representation as the error signal. Gain was updated after each item at test as follows:

$$g_{p\_new} = g_{p\_old} + \gamma * \Sigma_i \, (\delta_{ip} * net_{ip})$$

where g_{p_new} is the new gain for the hidden unit pool p (ranging over the weight hidden unit pool and the distance hidden unit pool), g_{p_old} is the old gain for the hidden pool p, δ_{ip} is the back-propagation delta term (Rumelhart, Hinton, & Williams, 1986) for unit i in pool p, net_{ip} is the net input to that unit, and γ is the learning rate parameter for gain adjustment. Activations of the units in the hidden pools were then computed for the next

item by applying the logistic function after scaling a unit's net input (with noise added, as discussed below) by g_{p_new}.

4.3 Noise

While the McClelland (1989, 1995) model was completely deterministic in its behavior during the test phase, it is clear that human behavior exhibits some variability. This variability is often thought of as one of the sources of innovation in behavior (as proposed, for example, by Siegler & Munakata, 1993). Many connectionist models capture variability by translating deterministic activations into probabilities at the response-selection stage (e.g., McClelland & Rumelhart, 1981). Subsequent research has indicated, however, that there can be problems with this approach: McClelland (1991) found that the policy used by McClelland and Rumelhart (1981) led to a poor fit to experimental data from experiments in which two independent cues to item identity were manipulated (e.g., Massaro & Cohen, 1983). In addition, this approach requires additional ad hoc assumptions to account for variability in reaction times, and there are further, more technical, difficulties (Ashby, 1982). A robust solution to these problems is provided by assuming that variability is actually intrinsic to processing (McClelland, 1991, 1993; Movellan and McClelland, 2001; Usher and McClelland, 2001), an idea with precedent in theoretical thinking about processing in neurons (Sejnowski, 1981). In keeping with the approach taken in McClelland (1991) and in Usher and McClelland (2001), a sample of normally distributed zero-mean Gaussian noise was added to a given unit's net input before its activation was calculated.

4.4 Parameters and simulation details

The gain learning rate parameter γ was set to 1 and the standard deviation of the noise was set to 0.1 in the reported simulations. Other values considered did not improve the fit to the data. All other parameters and training and testing procedures are the same as in the simulation above with the McClelland (1989) model. The results presented below come from simulation of 30 independent networks, tested at the end of each epoch of training, using epochs 5 to 60 in analysis. While attention may vary to some degree during the naturalistic experiences that are thought to give rise to the connection weights in the network, we treat naturalistic learning episodes as spaced far enough apart so that attention would revert to its default value between successive episodes. Accordingly, gain was fixed at 1 during training of all of the networks, and was reset to 1 at the beginning of each test session.

4.5 Results

The pretest, posttest, and divergence test distributions all showed bimodality and an inaccessible region (at least to a degree similar to that found by Jansen & van der Maas, 2001), which were especially pronounced in the posttest and divergence test distributions (see Figure 6). The modes in all three distributions were at the expected scores.[1] The pretest and posttest distributions were significantly different [χ^2 (5, N = 1680) = 28.6372, $p <$ 0.0001], with a trend from pretest to posttest similar to that found by Jansen and van der Maas (2001). We did a one-tailed sign test across networks to determine whether there was an overall increase in scores from pretest to posttest. The mean scores on the pretest and posttest were compared for each of the 30 networks, and the null hypothesis, that there is no tendency for an increase in scores from pretest to posttest, was rejected with $p < 0.0001$. The null hypothesis was also, and as definitively, rejected for the Jansen and van der Maas data.

[1] The pretest, posttest, and divergence distributions in the Jansen and van der Maas data are shifted toward lower scores compared to the model, indicating a difference in the overall distribution of abilities between the model and the participants tested in the Jansen and van der Maas experiment. Whereas the sampling in the network was uniform across epochs, the sampling of children was not completely uniform across ages. Since our focus is on patterns of transition, we did not adjust the distribution of epochs used in testing to more closely match those of children.

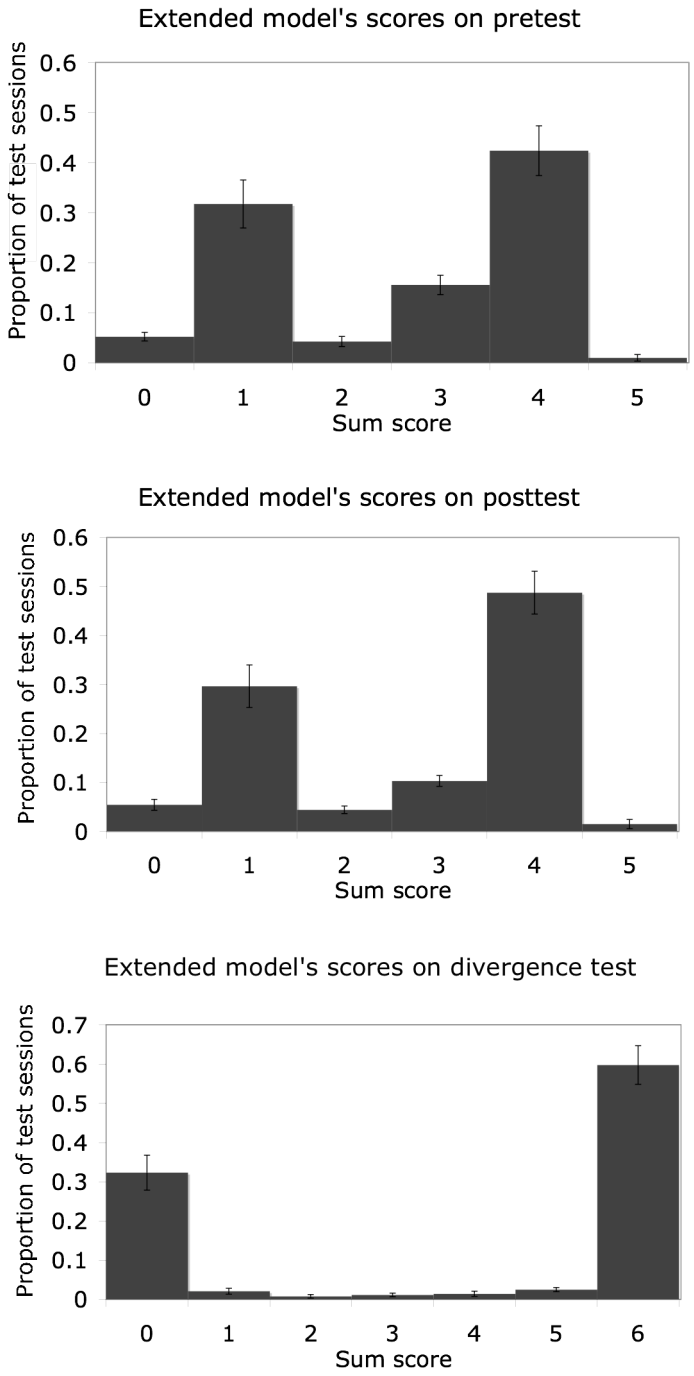


Figure 6. Results from an extension of McClelland's (1989) model with noise and gain. Distributions from 30 networks of pretest, posttest, and divergence test session scores. There were again 56 scores per network. Confidence interval calculations were done in the same way as in Figure 5.

We also did a more detailed analysis of the trends in the scores from pretest to posttest both for individual test sessions of the networks and for individual children. The purpose of this analysis was to determine whether the change from pretest to posttest involved a graded or sudden transition from Rule I to Rule II performance, both for children and for the networks. As shown in Table 1, Rule I or Rule II behavior was maintained consistently from pretest to posttest for 49.664% of the children and for 66.131% of the test sessions of all 30 networks. For both the children and the networks, where there was a pretest to posttest change, it was often from a score of one to a score of two or three, or from a score of two or three to a score of four. For example, of the children who had a score of one on the pretest and a higher score on the posttest, the score actually fell in the grey zone between one and four 57.409% of the time. For the model, the score fell in the grey zone 89.222% of the time. Overall, these data indicate that the change from pretest to posttest is often graded in children, but more discrete changes also occur. In the model, graded changes predominate, although there are some cases of an apparent "discrete stage" transition.[2]

[2] We explored using a higher level of noise in the model to see if this would produce more large changes in score from pretest to posttest. While this did produce greater variability in both the pretest and posttest scores, it did not result in more jumps from a score of one to a score of four.

Table 1

Percentages of Children and Model Test Sessions Displaying Each Combination of Pretest and Posttest Score

Children

Pretest score	Posttest score						
	0	1	2	3	4	5	*Pretest totals*
0	3.691	0.671	0.671	0.000	0.000	0.000	5.034
1	0.671	33.221	5.705	4.698	6.711	1.007	52.013
2	0.000	1.678	0.671	1.007	0.671	0.336	4.362
3	0.336	0.336	1.342	1.342	4.698	0.336	8.389
4	0.000	0.000	0.000	3.356	16.443	4.698	24.497
5	0.000	0.000	0.000	1.007	1.342	3.356	5.705
Posttest totals	4.698	35.906	8.389	11.409	29.866	9.732	

Model

Pretest score	Posttest score						
	0	1	2	3	4	5	*Pretest totals*
0	4.583	0.595	0.000	0.000	0.000	0.000	5.179
1	0.833	27.024	2.262	1.190	0.417	0.000	31.726
2	0.000	1.607	1.250	1.012	0.357	0.000	4.226
3	0.000	0.357	0.833	5.655	8.631	0.060	15.536
4	0.000	0.060	0.060	2.440	39.107	0.714	42.381
5	0.000	0.000	0.000	0.000	0.238	0.714	0.952
Posttest totals	5.417	29.643	4.405	10.298	48.750	1.488	

Note. Children with any missing values in the pretest or posttest were removed from analyses. $N = 298$ for children; $N = 1680$ for model.

The divergence hypothesis was again not supported by the network's results. There was no significant difference between the distribution of scores for items with one weight and the distribution of scores for items with five weights [$\chi^2 (3, N = 1680) = 0.9673, p = 0.8092$].

Table 2

Percentages of Children and Model Test Sessions Displaying Different Types of Patterns in Hysteresis Test

	Distance difference			
Type of pattern	First correct rising	Last correct falling	Children	Model
Rule I	-	-	31.847	29.464
Rule II	1	1	32.484	51.012
Hysteresis	5	4	0.000	-
Hysteresis	5	3	0.955	-
Hysteresis	5	2	0.318	-
Hysteresis	4	3	0.318	0.595
Hysteresis	4	2	0.000	0.417
Hysteresis	3	2	0.637	1.012
Maxwell	2	2	1.592	3.214
Maxwell	3	3	1.592	1.012
Maxwell	4	4	0.318	1.250
Maxwell	5	5	0.000	-
Sudden jump	2	1	4.140	4.881
Sudden jump	3	1	0.637	0.298
Sudden jump	4	1	1.592	0.000
Sudden jump	5	1	1.911	-
Residual	-	-	15.924	6.845
Missing Values	-	-	5.732	-

Note. The second column indicates the distance difference in the first half of the hysteresis items (as the distance difference rises) at which responses begin to be correct, and the third column indicates the last distance difference in the second half of the hysteresis items (as the distance difference falls) at which responses are still correct. N = 314 for children; N = 1680 for model. Dashes are given for untested model conditions (in the model the maximum distance difference is four).

Both the hysteresis and sudden jump catastrophe flags were found in the model's performance on the hysteresis item set. Of the 1680 total test sessions, 34 (2.024%) were hysteresis patterns, 87 (5.179%) were sudden jump patterns, and 92 (5.476%) were Maxwell patterns. In the control test, there was 1 (0.060%) hysteresis pattern, 7 (0.417%) sudden jump patterns, and 8 (0.476%) Maxwell patterns. The difference between the distributions on the two tests was highly significant [χ^2 (5, N = 1680) = 189.5052, $p <$ 0.0001].

Table 2 shows detailed data comparing the types of patterns displayed by the children and the model in the hysteresis test. For each pattern, we indicate the size of the distance difference for the first item correct in the sequence of items with rising distance difference and of the last item correct in the sequence of items with falling distance difference. Of particular interest is the tendency for the 'sudden jumps' displayed by both the model and the children to occur at small values of the distance difference. For both the children and the model, by far the most common sudden jump involves shifting from the incorrect 'balance' response when the distance difference is one to the correct distance-based response when the distance difference is two, and then persisting with this same response through the rest of the hysteresis test, including the final test item for which the distance difference is one. In the network, this kind of pattern arises when the network is already somewhat sensitive to distance difference, so that once the distance difference starts to grow, a shift to Rule II behavior occurs. Modest gain adjustment is then sufficient to produce a slight change in sensitivity to distance that looks like a 'sudden jump'. It is true that the children are more likely than the model to show larger

sudden jump patterns, but these cases are quite rare: cases in which the ‘sudden jump’ occurred at a distance difference of four or five only occurred in a total of 3.503% of children (11/314).

Similarly, when there is a hysteresis pattern, the extent of the hysteresis effect is often rather small, both for the children and for the network. Only seven children (2.229%) showed hysteresis effects, and of these, three involved a shift of only one step, three a shift of two steps, and one a shift of three steps. The model had about the same rate of hysteresis, but there was a larger sample of cases (34) due to the large number of model test sessions (1680). For the model, a shift of size one occurred in 79.397% of the cases, but shifts of size two were also observed (7 cases). Finally, it is worth noting once again the presence of Maxwell patterns in both the children and the network. Such patterns occurred in 3.503% of the children and 5.476% of model test sessions.

5. General Discussion

In two simulations, we have explored the ability of a continuous connectionist model to capture the patterns of change seen in the behavior of the 314 children tested in the balance scale task experiment of Jansen and van der Maas (2001). Even without any ability to change during testing, the model captured many features of the data, including two of the catastrophe flags they considered, as well as several signs of a degree of graded sensitivity to the distance dimension. These signs include partial inaccessibility—both children and the network produced scores in the inaccessible region some of the time—and patterns such as the Maxwell pattern on the hysteresis test. We then incorporated into the network the ability to modulate attention in a testing environment and introduced noise into processing, which has resulted in an extended version of the model that exhibits much more of the detailed structure of the data. Among these details is the presence of two additional catastrophe flags: hysteresis and sudden jump.

Clearly, our model in its original form was not sufficient to account for all of the data. Some mechanism for change in the absence of explicit feedback had to be added to address the fact that some children in the Jansen and van der Maas (2001) experiment progressed in their performance on balance scale distance problems during the course of the experiment. We relied on two factors that have been incorporated into other models: intrinsic noise in processing (McClelland, 1991; Usher and McClelland; 2001) and modulation of attention (Krushke, 1992; Krushke & Movellan, 1991). Incorporating these factors, and also using the network's own output to drive gain adjustment, were critical to allow the model to successfully account for these findings. Although the use of the network's own output to drive gain adjustment is a new assumption, it is similar to the

use of the network's own output to drive connection-based learning. Output-driven learning—often implemented by a "Hebbian" learning rule—may play a role in the ability to learn without external outcome feedback in other tasks. One such case arises in McCandliss et al. (2002), in which Japanese adults learned to discriminate exaggerated /r/ and /l/ sounds without feedback. This learning has now been modeled using a Hebbian, outcome-driven learning rule by Vallabha and McClelland (2007).

The issue before us now is this: should the transition from reliance only on weight in the balance scale task to a reliance on distance when the weight on the two sides of the balance scale is the same be viewed as a continuous or a discrete transition? Drawing on the cusp model, Jansen and van der Maas have used the presence of the four catastrophe flags as evidence that the transition between Rule I and Rule II behavior is discontinuous. The presence of hysteresis, in particular, has been considered sufficient evidence for discontinuity (Jansen & van der Maas, 2001, p. 457, Quinlan et al., 2007, p. 421, Raijmakers et al., 1996, p.105) and a means of distinguishing between acceleration and discontinuous change (Jansen & van der Maas, 2001, p. 452).

We have now presented a simulation of a model that exhibits these catastrophe flags. Are we forced to conclude that the model is exhibiting discontinuous change? While we are sympathetic to the cusp model's ability to capture discontinuities that can arise from an underlying continuous change, there are many reasons to think that our model is not really exhibiting a true discontinuity. In the model, the vast majority of the transitions that occurred from Rule I behavior were not to Rule II behavior. Though there was an overall trend for scores to improve from the pretest to the posttest, scores were not generally jumping from one to four from the pretest to the posttest. The presence of

hysteresis patterns in the hysteresis test also suggests a graded sensitivity to distance in the model, since the extent of the hysteresis effect was generally small. In addition, the sudden jumps that occurred in the hysteresis test tended to occur at low values of distance difference, as would be expected if an already moderate degree of sensitivity to distance difference underwent a slight increase. Based on the above, we conclude that the model is capable of (various degrees of) incremental change and that such incremental change can underlie the appearance of the four catastrophe flags seen in the Jansen and van der Maas data.

Given that the model could exhibit these flags through incremental change, what are we to think about their meaning when they occur in data from children? In large part, the children exhibited the same qualitative signs of continuity and graded sensitivity that the network did. As in the network, the majority of the transitions that occurred from pretest to posttest were not from Rule I to Rule II behavior. The children had a significant number of scores in the inaccessible region, suggesting again some graded sensitivity to the distance dimension. Like the network, a few children exhibited Maxwell patterns in the hysteresis test. Many children show 'sudden jumps' on the low end of the distance difference progression and small shifts in performance on the hysteresis test. All these are patterns consistent with graded sensitivity to distance difference and incremental change in that sensitivity during testing.

With these ideas in mind, it is worth discussing other perspectives and models of children's behavior on the balance scale task. First, in order to address patterns of inconsistent responding in the balance scale and other tasks, Siegler (1996) introduced the 'overlapping waves' model. According to this model and variants, discussed in

Siegler and Chen (2002) and Jansen and van der Maas (2002), children can switch back and forth between the use of different explicit rule representations (or 'strategies') in the transition between periods of consistent use of rules, with the probability of selecting each strategy gradually changing during the overlap period. While this is gradual change in some sense, it is not what we mean when we argue for continuity in transition. According to our approach, a child who responds 'balance' with a distance difference of one or two but who says that the side with the greater distance will go down for larger distance differences is not switching between rules in this explicit sense, but simply exhibiting graded sensitivity to the distance dimension. Although this child's behavior could be described as one of switching between one rule or strategy and another, some new mechanism must now be invoked by the rule theorist to explain just why and how a greater distance difference will trigger selection of a different rule.

The ACT-R model of the balance scale task from van Rijn, van Someren, and van der Maas (2003) uses the concept of saliency to explain how sensitivity to distance difference in periods of transition can cause selection of a new rule. When the distance difference becomes noticeable enough (as determined jointly by the magnitude of the distance difference and a saliency parameter), the distance 'property' is 'retrieved' in a search for possible properties on which the two sides of the balance scale may differ. This in turn triggers encoding of the fact that the distances actually differ. This approach shares with ours the idea that distance difference has a graded impact, but differs from it in that their model's retrieval of the distance property, and hence its use of distance information, is an all-or-none event. If activation of the distance property is sufficient to cross a threshold, distance information is used; otherwise it is not used.

Both our model and the model of van Rijn et al. can address the Maxwell, hysteresis, and sudden jump patterns seen in children's performance. The van Rijn et al.

model explains the delay and sudden jump patterns by an additional factor: an increase in baseline activation of the distance property that occurs as a consequence of retrieval of this property. If the increase in baseline activation is small, the Maxwell pattern occurs; if it is large, the sudden jump pattern occurs, and if it is intermediate, the delay pattern occurs. Again, the mechanism is similar to ours in that a prior use of distance information increases the tendency to use such information on subsequent problems.

Given that both models can explain all three patterns, is there any reason to prefer one account over the other? Perhaps both models will turn out to be able to account equally well for the Jansen and van der Maas data. However, it is worth noting that this remains to be seen for the model of van Rijn et al. These investigators did show that, with different parameter values, the Maxwell, hysteresis, and sudden jump patterns could all be produced. However they do not provide a quantitative comparison of their model's pattern of behavior to the patterns seen in the data from children's behavior. As a result, we do not yet know if the mechanisms in the model would provide a consistent account of the rate of Maxwell, sudden jump, and delay patterns while at the same time producing approximately the right distribution of pretest and posttest scores with approximately the right amount of change from the pretest to posttest. Thus, at this point, there is really no basis for knowing whether their model could match these patterns in the data as well as ours has done. We would certainly welcome a more detailed analysis to understand how well their model can account for the details of children's performance.

A further point in considering the two models is that in ours, an underlying mechanism is proposed that produces developmental differences in sensitivity to the weight and distance cues as a function of experience outside the laboratory testing situation. The van Rijn et al. model, in contrast, currently stipulates that at certain points

in development changes in sensitivity to weight and distance occur, allowing first Rule I and then Rule II performance. A strength of our approach, not shared by the van Rijn et al. model in its current form, is the fact that ours provides a mechanistic simulation of the underlying developmental change itself, rather than assuming that such change occurs in order to account for developmental transitions. We would also note that the gradual change in sensitivity to distance information in the model mirrors the age-dependent graded change in sensitivity to distance variation seen in the experiments of Wilkening and Anderson (1982, 1991), in which participants were allowed to make graded responses. To our knowledge the van Rijn et al. model has not yet been applied to this phenomenon, and new mechanisms for graded reliance on distance information may turn out to be necessary for such an application.

In the simulations presented, we have not addressed several of the important aspects of behavior beyond the Rule II pattern on the balance scale task. In particular, we have not considered Rule IV behavior or the nature of the transitions that occur in the grey zone between Rule II and Rule IV. We chose to focus on the data surrounding the transition from Rule I to Rule II since it is here that others have tended to see some of the clearest support for a discontinuous change in behavior. However, some comment on these later phases of development is in order.

When children have reached a point in development where they are sensitive to both the weight and distance dimensions and these dimensions are placed in conflict, intuition may well become an insufficient basis for response. In order to respond accurately in such cases, it may be necessary to employ a more explicit strategy. It has been previously argued that true Rule IV behavior that is often observed in older children

and adults may involve an explicit multiplication of weight and distance to calculate which side of the scale has greater torque (see McClelland 1989, 1995; McClelland & Jenkins, 1991, for discussion). Children (and adults) who appreciate that weight and distance are both important but who do not 'know' the torque rule may resort to other (imperfect) strategies, including guessing (as in Siegler's Rule III), and possibly the Buggy Rule and/or the Addition Rule (Ferretti et al., 1985; Normandeau et al., 1989). For the simple cases (weight, distance, and balance items), where the cues are not pitted against each other, we suggest that more implicit tendencies of the kind exhibited by our model may characterize the behavior of many, if not all, children. Previous work with our model also indicates that patterns of response choices on conflict problems similar to those produced by the strategies mentioned above—particularly, the Addition Rule or Buggy Rule—can also arise in our model. (Since the model exhibits the use of rules only on a descriptive level, it would account for the identical patterns of responses expected by use of the Addition Rule or Buggy Rule, but would not predict the RT effects that would be expected for the explicit use of either of these strategies.) To us there therefore remains considerable uncertainty about the degree to which performance in the grey zone between Rule II and Rule IV is based on implicit or explicit processes. It is possible that the best account will involve a mixture of explicit and implicit strategies.

Even in earlier phases of development, it is possible that explicit rules are used by children in some cases, and there are signs that for some children the transition from Rule I to Rule II behavior may be abrupt. While most children showed incremental change, there were also a few children who went from scores of one on the pretest to four on the posttest, and such transitions would be expected if children were using explicit rules.

Similarly, in the hysteresis test, a few children exhibited sudden jumps at large distance differences. We stress that such events were relatively rare even in children, and that even when they occur, caution about whether a truly discontinuous change has occurred may be in order. Though the model exhibits large transitions less frequently than children do, it does exhibit them sometimes, suggesting that a continuous mechanism may be at play even in cases where children's behavior shows an apparent sudden transition. It is also possible that there would be other ways to adjust a connectionist network (e.g., by making it recurrent rather than strictly feed-forward, or by introducing lateral inhibition) that would give rise to a greater degree of apparent discontinuity from a change process (such as connection or gain adjustment) that is underlyingly incremental in nature.

6. Conclusions

We have presented a connectionist model that accounts for many of the trends found in children's transition from Rule I to Rule II behavior on the balance scale task by Jansen and van der Maas (2001). A close look at this transition in children and in the model has shown that many of the details of the data seem to be more consistent with the continuous than the rule-based perspective, for many if not all children. Children go through periods of stable performance that can be described by certain qualitative properties; e.g., showing little or no sensitivity to an important cue such as distance while performing correctly with respect to the weight cue, or performing correctly on problems requiring use of both weight and distance cues. The transitions between these qualitative patterns are not always discontinuous, however, since intermediate behavior is often observed in the transitional periods. Describing performance in terms of rules can provide an approximate characterization of behavior, since children spend most of their developmental time in periods of apparent stasis between transitions. This relative accuracy of description, though, does not imply that the knowledge representation that causes this behavior is in the form of rules, and the observed behavior at transitions and our model's ability to account for it suggest that in fact this is often not the case.

The true picture likely involves a complex relationship between implicit and explicit levels of knowledge representation and the roles that they play in determining the way that children solve the items on the balance scale task. Perhaps children are forming explicit rules as approximate self-descriptions of their implicit knowledge or are integrating these different types of knowledge in some other way (see McClelland, 1995, for further discussion). In our perspective, we do not take the position that that there is no

such thing as discrete and explicit knowledge. Our view is simply that performance may often be based on implicit knowledge instead of, or in addition to, explicit rules or strategies.

The overall level of success of the presented model of the transition from Rule I to Rule II behavior and the clear presence of signs of graded sensitivity to distance in the data strongly suggest that a rule-based perspective (Siegler & Chen, 2002; Jansen & van der Maas, 2001, 2002; Quinlan et al., 2007) cannot be the complete account of developmental change on the balance scale task. While some of the cited papers have conceded that discrete rules are not always the whole story, their authors nevertheless persist in the view that a rule-based approach is fundamentally correct. Our findings challenge these researchers to show convincing evidence that explicit rules are ever needed to account for the transition from ignoring distance completely ('Rule I') to taking it into account when another stronger cue is in balance ('Rule II'). Though some of the children's transitions are of the kind expected from shifts between different rules, even these can sometimes be observed in our continuous model of this transition.

Several patterns of behavior that were thought to suggest rule use have been displayed by a model that does not explicitly represent rules, and observed behavior that deviates from the rule-based perspective has proven to fit naturally into the continuous perspective. Our findings suggest that models that stress underlying continuity in behavior have a significant role to play in the emerging picture of the mechanisms of developmental transitions.

Acknowledgements

I am tremendously grateful to Jay McClelland for his advice, encouragement, time, and immense insight throughout the course of this project. I feel indebted to him for helping me start on a career path where I can hopefully do much more of this kind of work. Working with him has been truly inspiring. I would like to thank the other members of the PDP lab and faculty in the psychology department, in particular Anthony Wagner, for their advice and support over the last few years. I would also like to thank my parents for their love and encouragement.

Appendix A

Comments on latent class analysis

Here we consider the latent class model and associated parameters from Table 5 of Jansen and van der Maas (2001), reprinted here as our Table 3. This table shows the probability of a child in the latent classes interpreted as Rule 0, Rule I, Rule II, or Rule III getting each item in the pretest and posttest correct. Matching superscripts indicate constraints imposed on parameter values assigned to particular problems in each latent class. These constraints specify that certain test items will share the same response tendency within the class. For example, within the 'Rule I' latent class, the probability of responding correctly is assumed to be the same for all three of the distance items.

Comparing the results of the pretest and posttest latent class analysis is complicated by the fact that the set of classes identified and the characteristics of each class are allowed to differ between the pretest and the posttest. To name three such differences: (1) Only three classes were identified in the pretest whereas four were identified in the posttest. (2) In the pretest, probability correct is constrained to be the same for all six problems in the Rule 0 latent class, while no such constraint is employed for the Rule 0 latent class in the posttest. (3) In other cases, the constraints imposed in the pretest and posttest for latent classes assigned to the same nominal rule are consistent, but response probabilities for a given type of item are allowed to differ between the pretest and posttest. For example, the probability of correct response to a distance item under the pretest 'Rule II' latent class is .90 while the corresponding probability for the posttest 'Rule II' latent class is .84.

These differences in the latent classes employed for the pretest and posttest make it difficult to compare the results of the two tests, since the same pattern of responding can be assigned to two different classes on the pretest and posttest. This different treatment of the same pattern of responding might explain the puzzling apparent decrease in the use of 'Rule II' from pretest to posttest, a trend inconsistent with the evidence of an increase in Rule II behavior shown in Figure 3. The first column of Table 4 shows the percentage of children in the pretest falling into each of the latent classes based on the probabilities given in Table 3 for pretest latent classes. The last column shows the percentage of children in the posttest falling into each of the latent classes based on the probabilities given for the posttest latent classes. The number of participants classified into the latent classes associated with Rule II decreases from the pretest to posttest.

An alternative approach to applying LCA to these data would be to treat performance in the pretest and posttest as drawn from the same set of classes, rather than from different sets of classes. Such an approach seems consistent with the viewpoint of Jansen and van der Maas, who view transitions in performance as reflecting a shift from one rule to another. We re-analyzed the data to see what it would look like to use consistent sets of classes across the pretest and posttest, and we present these results in Table 4: the second column categorizes posttest results using pretest latent class probabilities and the third column categorizes pretest results using posttest latent class probabilities. Comparing the first pair of columns, which now involve the same set of latent classes and class probabilities, we see the expected increase in Rule II use. The same holds for the second pair of columns. This analysis supports our point that different

decisions about how to constrain the application of LCA to the data can lead to different apparent patterns of change from pretest to posttest.

Table 3

Table 5 from Jansen and van der Maas (2001)

Estimated Parameters for Latent Class Models of the Pretest and the Posttest, Experiment 1

	p(Item = correct)						
p(l.c.)	Distance 1	Weight	Distance 2	CW	Distance 3	CD	Interpretation
			Pretest				
.52	.02[1]	.98[2]	.02[1]	1.00[3]	.02[1]	.00[3]	Rule I
.39	.90[4]	.98[2]	.90[4]	.75[5]	.90[4]	.25[5]	Rule II
.09	.16[6]	.16[6]	.16[6]	.16[6]	.16[6]	.16[6]	Rule 0
			Posttest				
.37	.03[1]	1.00	.03[1]	.97[2]	.03[1]	.03[2]	Rule I
.32	.84[3]	.95	.84[3]	.94[4]	.84[3]	.06[4]	Rule II
.09	.06	.13	.12	.04	.25	.22	Rule 0
.21	1.00	.91	1.00	.42	.95	.47	Rule III

Table 4

Percentages of children falling into each of the pretest-based and posttest-based latent classes when each set of classes is applied to performance at pretest and posttest

	Pretest-based classes		Posttest-based classes	
	at pretest	at postest	at pretest	at posttest
Rule I	52	37	52	37
Rule II	39	50	24	32
Rule III			15	21
Rule 0	9	12	10	9

Appendix B

Details of network testing items

			weights left		weights right	
item	part of test	item type	number	position	number	position
1	Practice	distance	2	5	2	3
2	Practice	conflict-distance	3	2	4	1
3	Practice	conflict-weight	1	5	2	4
4	Practice	weight	2	3	4	3
5	Pretest	distance	5	1	5	3
6	Pretest	weight	5	4	3	4
7	Pretest	distance	3	2	3	4
8	Pretest	conflict-weight	1	5	2	3
9	Pretest	distance	4	3	4	4
10	Pretest	conflict-distance	3	3	5	1
11	Hysteresis test	distance	5	1	5	2
12	Hysteresis test	distance	5	1	5	3
13	Hysteresis test	distance	5	1	5	4
14	Hysteresis test	distance	5	1	5	5
15	Hysteresis test	distance	5	1	5	4
16	Hysteresis test	distance	5	1	5	3
17	Hysteresis test	distance	5	1	5	2
18	Posttest	distance	5	3	5	1
19	Posttest	weight	3	4	5	4
20	Posttest	distance	3	4	3	2
21	Posttest	conflict-weight	2	3	1	5
22	Posttest	distance	4	4	4	3
23	Posttest	conflict-distance	5	1	3	3
24	Control test	distance	5	1	5	4
25	Control test	distance	5	1	5	4
26	Control test	distance	5	1	5	4
27	Control test	distance	5	1	5	4
28	Control test	distance	5	1	5	4
29	Control test	distance	5	1	5	4
30	Control test	distance	5	1	5	4
31	Divergence test	distance	1	3	1	1
32	Divergence test	distance	5	1	5	3
33	Divergence test	distance	1	2	1	4
34	Divergence test	distance	5	3	5	1
35	Divergence test	distance	5	2	5	4
36	Divergence test	distance	1	1	1	3

Modifications from the Jansen and van der Maas (2001) items are a change in the position of the weights to the left in the third practice item from 6 to 5, and two items deleted from each of the hysteresis test and control test to account for the network's maximal distance difference of four.

Appendix C

Explored Extensions

C.1 Learning during test

As mentioned in the section on Extensions to the Model, one way that the self-teaching signal can be used is in changing the network's connection weights during the test sessions using the network's categorized overt response as the basis of the error signal. In our investigations of this procedure, we discarded the weight changes at the end of each test session, so that a network's performance on a given test session would not not be affected by learning on earlier test sessions. This policy is appropriate in our case because we are modeling effects of experience during testing in children who are never tested twice. Rather than re-initialize each trained network from scratch after each test, we simply discarded any changes that occurred during testing before continuing with network training. Thus, at the beginning of each test session, the weights were saved, and after test they were restored to the saved copy before the next epoch of training.

Training the network during the test phase on its own responses did result in a significant number of patterns of hysteresis and sudden jump, but this policy for changing the network's weights caused the network's performance on the posttest to be worse than on the pretest. The reason for this is that the weight changes produced by a response favoring, say, the left side down tended to induce a position bias, favoring left side down responses to later items. This produced a disadvantage in the posttest, since the pretest distance items and hysteresis items favored one side of the scale while the posttest distance items favored the other. It is interesting that hysteresis and sudden jump patterns thought to indicate a phase change can coexist with a general decrease in level of

performance. However, the model's tendency to perform worse on the posttest than the pretest is not consistent with Jansen and van der Maas's (2001) findings, and whatever change is occurring in children's performance is unlikely to depend on a learned position bias.

C.2 Forcing Symmetry

As discussed above, the weight adjustment during test induced a side preference in the model which was not apparent in children's behavior, since in the specific materials used by Jansen and van der Mass, the pretest and hysteresis test all relied on distance items with the greater distance on the same side, and the posttest items used greater distance on the opposite side, and children performed better, not worse, on the posttest. We explored two different ways of eliminating this bias. Both of these approaches directly build into the model a constraint that children appear to follow in their learning; whether such a constraint could itself be learned is an open question. Since our focus was on continuity of stage transitions, the actual source of the symmetry constraint seemed a separate matter that we set aside for later consideration.

C.3 Forcing symmetry by symmetrizing the test set

Symmetry can be enforced by presenting the network with each item in the test set followed by the reflected version of that item (for example, one weight at a distance of five on the left and five weights at a distance of one on the right followed by five weights at a distance of one on the left and one weight at a distance of five on

the right). This eliminated the side preference, and allowed for positive transfer from pretest to posttest, and created a more pronounced inaccessible region in the pretest and posttest distributions, but was not enough to produce the degree of learning from pretest to posttest seen in the children.

C.4 Weight slaving

With the 4-input version of the architecture described below, weights that represent corresponding information for the two sides of the scale can be slaved together, treating the two hidden units responsible for each dimension as symmetrical so that one will represent greater weight or distance to the left, and the other will represent greater weight or distance to the right. This was done by averaging corresponding weights after any learning occurred (e.g., the weight from the input unit for distance on the left to the first hidden unit in the hidden layer for distance is averaged with the weight from the input unit for distance on the right to the second hidden unit in the same hidden layer). This had a similar affect on the results as the reflected test set had, and still did not produce a close enough correspondence to the learning trends seen in children.

It should be noted that the background training set used for the balance scale enforced symmetry only because there is in fact no statistical bias over the full set of training patterns favoring either the left or the right side to be the correct answer. Because the background training produces an essentially symmetrized network, the manipulation of gain at the hidden layers tends to preserve the symmetry; it simply scales up the magnitude of the influences the input units have on the hidden units.

C.5 4-Input Architecture

Another possible modification to the model is to use an architecture with only four, as opposed to twenty, input units. Instead of representing whether an individual weight is on the scale and whether a weight is at a particular distance with units for each piece of information that have activations of 0 or 1, we had four input units: one for weight on the left, one for weight on the right, one for distance on the left, and one for distance on the right (this was also done by Shultz, Mareschal, & Schmidt, 1994). Each of these units had activations of 0.2, 0.4, 0.6, 0.8, or 1, depending on whether there are 1, 2, 3, 4, or 5 weights at 1, 2, 3, 4, or 5 distances on each side. The use of distinct units to represent different amounts of weight and distance was discussed by McClelland (1995) as a shortcoming of the model because it does not allow the network to extrapolate beyond or interpolate within the range of values that it has experienced. On the other hand, using this ordered representation of the distance and weight encodes the structure of each dimension inherently so that the network does not do the work of learning these relationships.

The 4-input architecture with training at test produced a remarkably pronounced inaccessible region in the pretest, posttest, and divergence test (much more pronounced than the data from Jansen & van der Maas, 2001). Often there were no scores of two occurring at all in the pretest and posttest, though there were always some scores of three. The types of transitions in behavior from pretest to posttest were still not matching the children's behavior in these simulations. Several variants of this architecture were explored (including versions in which bias weights were eliminated, and in which symmetry was forced by weight averaging, as described above). Although some variants

produced results addressing many features of the Jansen and van der Maas (2001) data, none produced as good a fit as the gain manipulation.

References

Ashby, F. G. (1982). Deriving exact predictions from the cascade model. *Psychological Review*, *89*, 599-607.

Boom, J., Hoijtink, H., & Kunnen, S. (2001). Rules in the balance: Classes, strategies, or rules for the balance scale task. *Cognitive Development, 16,* 717–735.

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

Ferretti, R. P., & Butterfield, E. C. (1992). Intelligence-related differences in the learning, maintenance, and transfer of problem-solving strategies. *Intelligence, 16,* 207–223.

Ferretti, R. P., Butterfield, E. C., Cahn, A., & Kerkman, D. (1985). The classification of children's knowledge: Development on the balance-scale and inclined-plane tasks. *Journal of Experimental Child Psychology, 9,* 131–160.

Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence.* New York: Basic Books.

Jansen, B. R. J., & van der Maas, H. L. J. (1997). Statistical test of the rule assessment methodology by latent class analysis. *Developmental Review, 17*, 321–357.

Jansen, B. R. J., & van der Maas, H. L. J. (2001). Evidence for the phase transition from Rule I to Rule II on the balance scale task. *Developmental Review, 21*, 450-494.

Jansen, B. R. J., & van der Maas, H. L. J. (2002). The development of children's rule use on the balance scale task. *Journal of Experimental Child Psychology, 81*, 383-416.

Kerkman, D. D., & Wright, J. C. (1988). An exegesis of two theories of compensation development: Sequential decision theory and information integration theory. *Developmental Review, 8,* 323–360.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review, 99*, 22-44.

Kruschke, J. K., & Movellan, J. R. (1991). Benefits of gain: Speeded learning and minimal hidden layers in back-propagation networks. *IEEE Transactions on Systems, Man and Cybernetics, 21*, 273-280.

Massaro, D. W., & Cohen, M. M. (1983). Phonological constraints in speech perception. *Perception & Psychophysics*, *34*, 338-348.

McCandliss, B. D., Fiez, J. A., Protopapas, A., Conway, M., & McClelland, J. L. (2002). Success and failure in teaching the [r]-[l] contrast to Japanese adults: Predictions of a Hebbian model of plasticity and stabilization in spoken language perception. *Cognitive, Affective and Behavioral Neuroscience. 2:2*. 89-108.

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

McClelland, J. L. (1991). Stochastic interactive processes and the effect of context on perception. *Cognitive Psychology, 23*, 1-44.

McClelland, J. L. (1993). Toward a theory of information processing in graded, random, interactive networks. In D.E. Meyer & S. Kornblum (Eds.), *Attention & Performance XIV: Synergies in experimental psychology, artificial intelligence and cognitive neuroscience (*pp. 655-668). Cambridge, MA: MIT Press.

McClelland, J. L. (1995). A connectionist perspective on knowledge and development. In T. Simon & G. Halford (Eds.), *Developing cognitive*

competence: New approaches to process modeling (pp. 157-204). Mahwah, NJ: Lawrence Erlbaum Associates.

McClelland, J. L., & Jenkins, E. (1991). Nature, nurture, and connections: Implications of connectionist models for cognitive development. In K. Van Lehn (Ed.), *Architectures for intelligence* (pp. 41-73). Hillsdale, NJ: Lawrence Erlbaum Associates.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: Part 1. An Account of basic findings. *Psychological Review*, *88,* 375-407.

Movellan, J.R., & McClelland, J. L. (2001). The Morton-Massaro Law of Information Integration: Implications for Models of Perception. *Psychological Review*, *108*, 113-148.

Normandeau, S., Larivee, S., Roulin, J. L., & Longeot, F. (1989). The balance-scale dilemma: Either the subject or the experimenter muddles through. *Journal of Genetic Psychology*, *150,* 237–250.

Piaget, P., & Inhelder, B. (1969). The psychology of the child. London: Routledge.

Quinlan, P. T., van der Maas, H. L. J., Jansen, B. R. J., Booij, O., & Rendell, M. (2007). Re-thinking stages of cognitive development: An appraisal of connectionist models of the balance scale task. *Cognition*, *103*, 413-459.

Raijmakers, M. E. J., van Koten, S., & Molenaar, P. C. M. (1996). On the validity of simulating stagewise development by means of PDP networks: Application of catastrophe analysis and experimental test of rule-like network performance. *Cognitive Science, 20*, 101-136.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal

representations by error propagation. In J. L. McClelland, D. E. Rumelhart, & the PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition: Vol. I* (pp. 318-362). Cambridge, MA: MIT Press.

Sejnowski, T. J. (1981). Skeleton filters in the brain. In G. E. Hinton and J. A. Anderson (Eds.), *Parallel models of associative memory* (pp. 189-212). Hillsdale, NJ: LEA Associates.

Shultz, T. R., Mareschal, D., & Schmidt, W. (1994). Modeling cognitive development on balance scale phenomena. *Machine Learning, 16,* 57–86.

Shultz, T. R., & Takane, Y. (2007). Rule following and rule use in the balance-scale task. *Cognition, 103,* 460-472.

Siegler, R. S. (1976). Three aspects of cognitive development. *Cognitive Psychology, 8,* 481-520.

Siegler, R. S. (1981). Developmental sequences within and between concepts. *Monographs of the Society for Research in Child Development, 46*(2), 1-74.

Siegler, R. S. (1996). *Emerging minds. The process of change in children's thinking*. New York: Oxford Univ. Press.

Siegler, R. S., & Chen, Z. (1998). Developmental differences in rule learning: a microgenetic analysis. *Cognitive Psychology, 36,* 273–310.

Siegler, R. S., & Chen, Z. (2002). Development of rules and strategies: balancing the old and the new. *Journal of Experimental Child Psychology, 81*, 446-457.

Siegler, R. S., & Munakata, Y. (1993). Beyond the immaculate transition: Advances in understanding developmental change. *SRCD Newsletter, Winter*, pp. 3, 10, 11, 13.

Usher, M., & McClelland, J. L. (2001). On the time course of perceptual choice: The leaky competing accumulator model. *Psychological Review*, *108*, 550-592.

Vallabha, G. K., & McClelland, J. L. (2007). Success and failure of new speech category learning in adulthood: Consequences of learned Hebbian attractors in topographic maps. *Cognitive, Affective and Behavioral Neuroscience, 7,* 53-73.

Van der Maas, H. L. J., & Jansen, B. R. J. (2003). What response times tell of children's behavior on the balance scale task. *Journal of Experimental Child Psychology, 85,* 141–177.

Van der Maas, H. L. J., & Molenaar, P. C. M. (1992). Stagewise cognitive development: An application of catastrophe theory. *Psychological Review, 99,* 395-417.

Van der Maas, H. L. J., Quinlan, P. T., & Jansen, B. R. J. (2007). Towards better computational models of the balance scale task: A reply to Shultz and Takane. *Cognition, 103,* 473-479.

Van Maanen, L., Been, P., & Sitjsma, K. (1989). The linear logistic test model and heterogeneity of cognitive strategies. In E. E. Rosram (Ed.), Mathematical psychology in progression (pp. 267–287). Berlin: Springer-Verlag.

Van Rijn, H., van Someren, M., & van der Maas, H. (2003). Modeling developmental transitions on the balance scale task. *Cognitive Science, 27*, 227-257.

Wilkening, F., & Anderson, N. H. (1982). Comparison of the two rule-assessment methodologies for studying cognitive development and structure. *Psychological Bulletin,* 92, 215-237.

Wilkening, F., & Anderson, N. H. (1991). Representation and diagnosis of knowledge structures in developmental psychology. In N. H. Anderson (Ed.),

Contributions to information theory: Vol. III Developmental (pp. 45-80).

Hillsdale, NJ: Lawrence Erlbaum Associates.