Machine Learning, ??, 1–30 (199?)

A Decision-Tree Model of Balance Scale Development

WILLIAM C. SCHMIDT wcs@is.dal.ca
Department of Psychology, Dalhousie University, Halifax, Nova Scotia, Canada B3H 4J1

CHARLES X. LING ling@csd.uwo.ca
Department of Computer Science, The University of Western Ontario, London, Ontario, Canada N6A 5B7

Editor: Michael J. Pazzani

Abstract. We present an alternative model of human cognitive development on the balance scale task. Study of this task has inspired a wide range of human and computational work. The task requires that children predict the outcome of placing a discrete number of weights at various distances on either side of a fulcrum. Our model, which features the symbolic learning algorithm C4.5 as a transition mechanism, exhibits regularities found in the human data including orderly stage progression, U-shaped development, and the torque difference effect. Unlike previous successful models of the task, the current model uses a single free parameter, is not restricted in the size of the balance scale that it can accommodate, and does not require the assumption of a highly structured output representation or a training environment biased towards weight or distance information. The model makes a number of predictions differing from those of previous computational efforts.

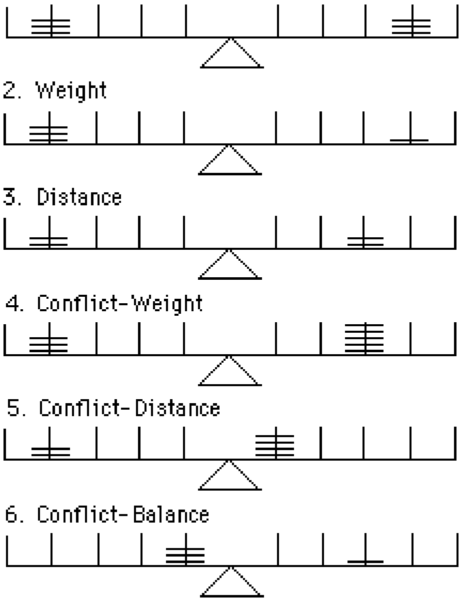
Keywords: Cognitive development, balance scale, C4.5, symbolic learning algorithms.

1. Introduction

Within the past decade a number of symbolic and connectionist learning methods have been applied to cognitive development's balance scale task. The aim of this body of research has been to investigate the use of machine learning methods as models of developmental transition, to explore the range of assumptions under which psychologically accurate models of the task can be achieved, and, most important, to assemble predictions about the task and the changes that children's thinking undergoes during the course of development.

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

The psychological task requires the integration of the dimensions of weight and distance through the course of development. Perfect performance on this task can be achieved by computing torques for both the left and right arms by multiplying



Problem Type	Stage 1	Stage 2	Stage 3	Stage 4
1. Balance	100	100	100	100
2. Weight	100	100	100	100
3. Distance	0	100	100	100
4. Conflict-Weight	100	100	33	100
5. Conflict-Distance	0	0	33	100
6. Conflict-Balance	0	0	33	100

Figure 1. Predictions of percent problems correct for children under different stages.

weight by distance, and the side with the largest torque goes down. If torques are equal, then the scale will balance.

Siegler (1976, 1981) has partitioned the set of possible balance scale problems into the six sets of distinct problem types shown in Figure 1.[1] Performance on the different problem types is used to gage the level of expertise that the child has acquired and to gain insight into the types of information that the child uses to solve balance scale problems.

The first three types of problems are referred to as *simple* problems since the dimension of greater magnitude determines the outcome. *Balance* problems have equal numbers of weights placed at equal distances from the fulcrum resulting in the scale balancing. Distances on either side of the fulcrum are equal for *weight* problems, resulting in the dropping of the side with more weight. For *distance* problems, the arm with weights placed at a greater distance goes down since the two sides have equal weights.

The final three problem types are referred to as *conflict* problems since the cue of weight conflicts with the cue of distance. These problems have greater weight on one arm and greater distance on the other, with no simple way of determining the outcome. The side with the greater weight or distance drops respectively in *conflict-weight* and *conflict-distance* problems, while the scale balances for *conflict-balance* problems.

Siegler (1976, 1981) has reported that children's performance on the balance scale task progresses through four distinct stages. In stage 1, children use only weight information to determine if the scale will balance. In stage 2, children emphasize weight information but use distance if weights on either side of the fulcrum are equal. In stage 3 both weight and distance information is utilized for simple problems, but children seem to respond indecisively when one arm has greater weight and the other greater distance. By stage 4, there is a correct integration of weight and distance information resulting in the near flawless performance of the task. Each of these rules makes specific predictions about the pattern of errors that can be expected to occur across the different problem types. Figure 1 presents the predicted percentages of correct responses, broken down by problem type, for each of these four stages.

While orderly stage progression constitutes a major regularity of balance scale development, a second regularity can be observed by examining the predicted pattern of errors in Figure 1 for conflict-weight problems. In stages 1 and 2, children answer these problems correctly since they generally rely on the weight dimension to determine their response. In stage 3 however, when weight and distance cues are in conflict, children often perform poorly on the same type of problems that they had previously answered correctly. This situation is rectified by stage 4 however, at which point correct answers re-occur. This trend is referred to in the developmental literature as U-shaped development, reflecting the pattern of the longitudinal plot of performance. The U is created by initially good performance followed by a regression and a later recovery, on conflict-weight problems.

A third major balance scale regularity was reported by Ferretti and Butterfield (1986). These researchers discovered that the rule classifications of many children systematically varied when assessed with different sets of testing problems drawn from the theoretically equivalent problem types. It was discovered that children's judgements about problems with a greater absolute difference in the amount of torque between the two arms (*torque difference*) were more often correct than similar types of problems with smaller torque differences. As a result, children assessed with problems using larger torque differences are classified at a higher developmental stage by Siegler's rule assessment procedure than if they were assessed with smaller torque difference problems. This last phenomenon is dubbed the *torque difference effect* (TDE).

2. Previous Models of Balance Scale Development

There have been three symbolic models and three connectionist models of the balance scale task published to date. The first symbolic approach modeled each of the four stages of balance scale development as a separate set of production system rules (Klahr & Siegler, 1978). Later, Sage and Langley (1983; Langley, 1987) expanded on Klahr and Siegler's findings by proposing a possible transition mechanism. The third symbolic learning attempt implemented a balance scale model using the Soar architecture (see Newell, 1990 for an in-depth overview of Soar's learning methods) as a transition mechanism.

The first connectionist balance scale model was created by McClelland (1989) using the back-propagation learning algorithm as a transition mechanism. Shultz and Schmidt (1991; Shultz, Mareschal, & Schmidt, 1994) later reported on the use of the cascade-correlation learning architecture to create a model making similar assumptions. Finally, Shultz, Schmidt, Buckingham, & Mareschal (1995) reported on a second cascade-correlation model of the balance scale which undertook a quite different set of assumptions. We will now examine each of these models in turn, pointing out some strengths and weaknesses.

2.1. Symbolic Models

The Klahr and Siegler (1978) work successfully captured performance at each stage of balance scale development as a separate, unchanging set of production rules. No attempt was made at specifying a transition mechanism, so it is not really applicable to question whether U-shaped development on conflict-weight problems occurred. Finally, the rules appear to be incapable of producing a TDE since they did not differentiate the magnitude of the weight or distance in a given balance scale problem.

Sage and Langley (1983; Langley, 1987) expanded upon Klahr and Siegler's (1978) approach by adding a transition mechanism. Their model started with a set of rules that made random predictions, but which learned to improve. The transition mechanism was a discrimination process that searched for differences between cases in which correct predictions were made and cases in which errors were made.

Sage and Langley (1983) reported that the system managed to pass through stages "similar" to stages 1, 2 and 3, and while the system learned rules similar to those specified by Klahr and Siegler (1978), it also learned other rules which conflicted with the Klahr and Siegler results. In particular, the model acquired rules in which distance was the principle dimension at the same time that it induced rules which placed weight in a dominant role. Unfortunately, the only rule diagnosis metric mentioned by the authors was the percentage of correct predictions made by the model based on its responses to a small number of selected problems. It is unclear from such an evaluation how strictly the model cohered to the stages observed in the human literature.

The final symbolic balance scale model to date was reported by Newell (1990) using the Soar architecture. The model was not assessed in a rigorous fashion, but reportedly proceeded to learn to correctly respond to five balance scale problems, transpiring through stages 1, 2 and 3 in the process. There was no comparison of the model's output to the human data, except for an overall qualitative judgement of stage transition. The psychological realism of the Soar model is questionable. The model learned from very few instances and an entire stage often consisted of witnessing only a single exemplar. Furthermore, it is unclear how dependent the Soar model was on observing specific problems in a certain order. The model was not examined for the TDE, and it is unclear how it could have captured this regularity.

In summary, the symbolic models of transition do not seem to have fared well on this task. None were successful in acquiring a stage 4 level of performance, it is questionable whether they demonstrated orderly rule progression or U-shaped development on conflict-weight problems, and it is not apparent whether or how these models might have captured the TDE. Perhaps this judgement would change if a more rigorous assessment procedure were used to evaluate the models' performances; however all indications are that these models would still fare poorly.

2.2. Connectionist Models

Our review of balance scale models now turns to connectionist accounts of the task. These approaches have performed much better and have been assessed in a much more thorough fashion than their symbolic counterparts. McClelland (1989) reported on the creation of a connectionist model of the balance scale task with five pegs and five weights per arm which used the back-propagation learning algorithm as a transition mechanism.

McClelland's model was characterized by two major assumptions. The architectural assumption required both a highly structured output representation and the separate processing of the weight and distance dimensions. The network had two output units with activation solely in a single unit representing left or right side down; if both units were active, the output was considered to be a balance response. In addition to structuring the output representation, the network topology enforced the separate processing of weight and distance information by feeding these dimensions into distinct sets of hidden units.

The second major assumption, the environmental assumption, had two components. The first component assumed that the children's environment does not continually present them with all of the possible balance scale problems, but that a randomly chosen and variable subset of these problems is required to succeed in learning on the balance scale task. The second component assumed that children gain more experience early on with weight information than they do with distance information. In particular, the model assumed that the child's environment is biased so that information from the weight dimension is more frequently available for predicting the problem's outcome than is information from the distance dimension.

The environmental assumptions were implemented by randomly selecting 100 patterns for training the network each epoch (an epoch is a single pass through the training instances for learning) and by supplementing the set of all possible training problems with a bias towards problems where distance on either side of the balance scale is equal (simple weight problems). This bias results in presenting the network with greater variation on the weight dimension making it easier for the network to learn weight than distance information.

McClelland's model successfully captured many of the details found in the human balance scale literature, including orderly stage progression, longitudinal U-shaped learning on conflict-weight problems, and the TDE (Schmidt & Shultz, 1991). The model failed to achieve a consistent level of stage 4 performance, and it also exhibited a response bias towards one arm of the balance scale. This bias resulted from the differential reduction of error on the two sides of the network caused by the network topology's incongruous treatment of weight and distance information. McClelland's model was the first balance scale model subjected to Siegler's rigorous rule assessment methodology.

Shultz and Schmidt (1991; Shultz et al., 1994) produced a second five-peg and five-weight connectionist model of the balance scale task using the cascade-correlation learning architecture. These researchers were interested in examining whether a generative connectionist algorithm, one that determines its own network topology, would be capable of producing a psychologically realistic model without making as many assumptions as McClelland's (1989) model did. This cascade-correlation model followed McClelland's lead by including a strong bias of equal distance problems in the training set and by training on a limited number of exemplars. The model began with a corpus of 100 randomly drawn training examples and gradually increased the base of instances from which the network learned each epoch. This model removed the architectural assumption of segregated weight and distance information processing that McClelland's back-propagation model had required while retaining an environmental assumption similar to McClelland's. The initial topology of the network consisted of four input units and two output units. Two input units for each side of the balance scale received integer values representing the weight and distance values of the problem, and the output units required interpretation in a manner similar to McClelland's model.

The cascade-correlation model demonstrated orderly stage progression, U-shaped development on conflict-weight problems and captured the TDE. In addition, the model successfully learned to a stage 4 level of performance without requiring an explicit rule system and did not exhibit a response bias (Shultz et al., 1994). This connectionist method also withstood the rigorous assessment methods applied to the human data.

A final connectionist model of balance scale development also used cascade-correlation and abandoned the assumptions of a biased or limited training environment, and the need to segregate weight and distance information processing in the network topology (Shultz et al., 1995). The topology of this model initially consisted of 10 input units and a pair of output units. The weight dimension of balance scale problems was encoded in a localist fashion (one unit representing each of the five possible weight values for each side of the balance scale) and the distance dimension was represented by turning on the input units proportionally to the magnitude of the distance associated with a given weight unit. The output units required interpretation as did previous connectionist balance scale simulations.

The Shultz et al. (1995) model prestructured the internal state of the network making an assumption that learning starts off from a position in the space of all possible sets of network connection weights which results in the preferential treatment of weight over distance information. After placing the network in such a position, the model was exposed to the full set of training exemplars and all 4 stages were achieved in an orderly fashion, as was U-shaped development on conflict-weight problems. Like the other connectionist models, the prestructured weight dimension model also exhibited the TDE and withstood rigorous assessment methods.

From this brief summary of balance scale models it would appear that symbolic transition mechanisms have little to offer cognitive development. We will demonstrate the appropriateness of symbolic learning methods for modeling cognitive development by presenting an alternative computational model of balance scale development using Quinlan's (1993) symbolic C4.5 machine learning algorithm as a transition mechanism. This C4.5 model provides not only a method for exploring developmental phenomena, but introduces a new set of assumptions capable of generating a realistic model of the balance scale task. These assumptions, coupled with the learning method, make a number of predictions about the task, and the changes that children's thinking undergoes during the course of development.

Ours is not the first developmental learning model to use C4.5. Ling and Marinov (1993) created a decision tree model of past tense language acquisition which provided a good fit to the human data while making only a small number of assumptions. In addition, the rules which were induced provided an interesting implementation of some proposals which had been prescribed by linguistic theory. It was in part the success of this model which motivated us to apply C4.5 to the current development problem.

3. C4.5 and Developmental Modeling

Since C4.5 will act as our transition mechanism, we now devote some space to describing its workings. We will first introduce C4.5 in a very general manner, and then examine in some detail how this algorithm can easily be applied to generating developmental models.

3.1. C4.5 — A Symbolic Classification System

We used Quinlan's (1993) C4.5 to act as a transition mechanism in our model. This general purpose classification system generates a simple (and small) decision tree capable of classifying data which vary along a number of specified attributes. C4.5 is a descendent of ID3 (Quinlan, 1986) which in turn found its origins in Hunt, Marin, and Stone's (1966) CLS decision tree induction system. Like back-propagation and cascade-correlation, C4.5 is a supervised learning algorithm.

Given a set of training examples which vary along a set of attributes, C4.5 extracts rule-like regularity from the examples and builds a decision tree which will classify the examples with some degree of tolerated error. For a given example, the nodes of the decision tree are used to test attribute values. The branch that is followed from a given node depends upon the outcome of the test. Each leaf of the tree is labeled by one of the possible classifications. If one traverses the path from the tree's root to a leaf according to the values that a problem possesses, then the classification found at the leaf is the problem's predicted classification.

C4.5 determines how to go about constructing a decision (sub)tree by computing a value called the *information gain ratio* (IGR) for each of the possible attributes which could potentially be used to partition the data. The IGR is a heuristic method that evaluates an attribute's ability to reduce randomness in unclassified examples. The attribute with the greatest IGR is chosen as the root of a subtree. Although IGR is a metric that makes a local decision as opposed to a globally optimal decision, the attribute selected is often the most discriminative. This method of building subtrees is applied recursively until the resulting tree fully classifies all of the training examples.

3.2. Modeling Cognitive Development with C4.5

To implement the transition component of the current balance scale model, the number of cases which were required to merit a subtree branching operation was varied with time. This manipulation resulted in the gradual emergence of an increasingly discerning decision tree. By assuming that what develops in children is an ability to assimilate more and more information over time, a series of decision trees can be constructed, each of which builds on its predecessor, and which can successively cover more and more of the data.

Given a set of training examples, there are several possibilities of generating a series of increasingly discerning decision trees. One obvious approach is to limit the depth of the decision trees, and to gradually increase the depth limit. However, this would produce trees with uniform depth, and the error rates in different leaves are uneven. Another approach is to limit the number (or percent) of errors that leaves are allowed to have, and to gradually decrease that limit. In this case, errors at different leaves would be uniform, and parts of trees with higher errors would be expanded more deeply. To model the balance scale task, we used a C4.5 parameter which has a similar effect.

C4.5 provides a user specified parameter, m, which during decision tree construction controls the minimum number of examples that branches of a decision node should classify. More specifically, for an attribute to be used as a decision node, it must have at least two branches, and every branch must classify m or more examples. If an attribute has only one branch that classifies m or more examples and all other branches account for less than m examples, then a split using this attribute is not considered useful (since all other branches account only for a minute number ($< m$) of examples), and a partitioning of the training data based on this attribute does not occur even if its IGR is highest.

Applying C4.5 with large values of m yields smaller decision trees since there may be few attributes that qualify to act as decision nodes. In such cases, the tree terminates with leaves whose class is the majority class of the examples falling into that node. The decision tree at this stage has a large error rate for the training examples. However, since C4.5 always chooses the most discriminant attribute as the root of (sub)trees, such small decision trees are, in some sense, the *best* small decision trees one could have to get as much regularity out of the training examples as possible. As m decreases, more nodes qualify to be split, more regularity in the training set is captured, and deeper trees are built. Therefore, the larger the m, the smaller the tree, and the less adequately the tree can accommodate the training examples; as m decreases, the promising branches are expanded to deeper levels and more training examples are classified correctly. If m is set to 1, as long as the training data are not conflicting (no two examples with the same attribute values but different classes exist), a decision tree that perfectly classifies every example in the training set results.

With respect to using C4.5 to model real world systems, if the target of the C4.5 model is limited in capacity (incapable of accommodating certain problems in order to attain task perfection) then C4.5 should be applied with a large m, producing the best small decision trees. As m decreases, more discriminating splits are added, resulting in larger, more discriminating decision trees being generated. The change in m corresponds to an increase in capacity; that is, an increase in mental ability. At this time, we make no claims about what sort of mechanism in the human learner m might correspond to. We simply wish to emphasize that as m varies, so too does the capacity of the resulting decision tree.

Based on this quality of m (or other tree-growing control parameters discussed earlier), we propose the following hypothesis for modeling development on the balance scale task (and in general, other tasks as well) using C4.5. Early in development, children have limited mental abilities. Their poor performance can be modeled with a large m value in C4.5. Performance and capacity improvement can be modeled by the gradual decrease of the m parameter in C4.5. The following simulations show that we can demonstrate major phenomena during children's development of the balance scale learning process through the series of decision trees with decreasing m in C4.5. This provides strong evidence that by decreasing m, C4.5 can provide a good developmental model.

4. A Decision-Tree Balance Scale Model

As mentioned earlier, there are three major phenomena that any model of the balance scale should aim to capture: systematic stage transition, U-shaped development on conflict-weight problems, and the TDE. In this section, we present two balance scale simulations. The first simulation shows orderly stage progression and U-shaped development on conflict-weight problems. The second simulation, which expands on the first, examines whether a symbolic learning algorithm can exhibit the TDE at all points in development. Following the model's presentation, we discuss a number of new predictions that the model makes about the psychological balance scale data.

4.1. Simulation 1: The Basic Model

We chose to simulate a five-peg, five-weight version of the balance scale in order to make our results directly comparable to those of the more successful connectionist balance scale models discussed previously. However, it should be noted that there is no reason why our model could not be successful with balance scale problem sets consisting of larger or smaller weights and distances. A discussion of this statement will appear later in this paper.

For a five-weight, five-peg problem set there are 625 unique problems (5 different weight values at each of 5 different distances on each balance scale arm, fully crossed between the two arms). For purposes of learning with C4.5, these problems need to be represented in terms of a set of values on attributes with an associated classification.

Regardless of the learning algorithm that one adopts (connectionist or symbolic), the choice of attributes to use is crucial if the model's output is to match the human data. More important, the set of attributes which yield a successful model make predictions about the types of information that humans may use, or are sensitive to, during development. The set of attributes that yield a successful model provide, at minimum, an existence proof about the types of information primitives that can be used to solve the problem in a developmental sequence similar to that exhibited in the human target.

Although we experimented with a number of different attribute sets, we found that very few led to a successful model of the human data. For instance, presenting C4.5 with simply the magnitudes of weights and distances for each side of the balance scale and a set of training examples with 90% of the problems of the simple weight type, as Shultz et al. (1994) did, led only to stage 3 and 4 performance. Similar results ensued using localist encoding of the inputs, and a trial corpus composed of 66% simple weight problems, as McClelland (1989) did. The assumptions underlying successful connectionist implementations simply did not yield a successful C4.5 implementation.

We present the set of attributes which we feel are crucial to our model's success, and which should be taken as predictions about the sorts of information that humans use during reasoning. This is a common approach to modeling in a domain where the object of the model is not fully understood. It should be noted that presented with just the basic balance scale problems, both connectionist and symbolic algorithms fail to yield realistic simulations. Both result in behavior classified at stages 3 and 4 only. By biasing the weight dimension and restricting the size of the training set, connectionist systems can model the human data. As we will show, decision tree systems can capture the human regularities if provided with redundant information that allows the learning algorithm to partition the data in a number of different ways.

Our model presents balance scale information to C4.5 using a set of seven attributes. Four of these attributes directly refer to the values on the weight and distance dimensions of the balance scale. The remaining three attributes present summary information about the problem. This summary information may be used as primitives in children's reasoning about balance scale problems. Siegler's (1976) work suggests that children reason with information about which side of the balance scale has the greatest weight or distance, and whether the sides of the balance scale are equivalent for a given dimension. Because of Siegler's success at capturing children's behavior with these primitives, we presented C4.5 with this sort of information.

The first model attribute concerns whether the problem presents an equal number of weights at equal distances on either side of the fulcrum, and can take values of yes or no. The inclusion of this attribute is based on the salience of simple balance problems. Such problems are the only examples which are wholly symmetrical, making them perceptually salient. In addition, such problems are conspicuous for children because they have the often intriguing outcome of balancing. Johnson (1987) outlines an extensive theory of development in which the child's schemata of balance plays a recurring and central role in development. Simple balance problems possess a favored status because they are the first and most simple cases that can be accommodated. Johnson's model, coupled with the fact that children perform perfectly on simple balance problems from the time that they are able to execute the balance scale task, suggests that simple balance problems play an important role in children's reasoning. Without this attribute information, pilot runs failed to acquire the balance concept early in development, and consequently failed to be classified at the earliest stages of performance.

The second and third attributes of the model concern which side of the scale has greater weight and distance respectively, and each takes on one of three values: left arm, neither arm, or right arm. These attributes suggest that human subjects might compare the two sides of the balance scale along either the weight or distance dimension. Siegler's rule models directly incorporate such information. Making this information primitively available to the learning algorithm presents it with the opportunity to capitalize on any informational value that side information may have for predicting problem outcomes. It should be noted that for children, this information can be immediately determined from the visual input (as can the simple balance attribute). By placing the weight attribute first, we can set up a situation where, faced with equally informative weight and distance information, the learning algorithm will select to rely on the first attribute type it encounters. This order effect is equivalent to assuming that children's development internally relies on information from one dimension over the other.

The fourth through seventh attributes are the actual number of weights and distances on either balance scale arm, and each of these is declared to be a continuous attribute taking on integer values ranging from 1 to 5. The inclusion of these attributes reflects that humans have such information readily accessible to them when confronted with balance scale problems. It should be noted that the model works equally well if these are treated as discrete attributes, however, such a representation prevents the induced decision trees from processing input values other than those in the training set.

Like all psychological models, our model necessarily makes a set of assumptions about the training environment and the information that is made available to children. Unlike many previous attempts, our model assumes that there is no explicit environmental bias for or against either the weight or distance dimension. It does, however, assume that simple balance problems, where an equal number of weights occur at an equal distance from either side of the fulcrum, are particularly salient for the purposes of children's learning. This hypothesis was reflected earlier by including this information as a separate attribute. We also implemented this assumption by including three times as many simple balance problems as naturally occurs within the problem set. Since there are only 25 simple balance problems (4% of the training examples) this amounted to the addition of an extra 50 problems, for a total of 675 training instances (making simple balance problems compose 12% of the training set). We feel that biasing simple balance problems is reasonable due to their high degree of salience. Balancing is an important part of any child's learning (see Johnson, 1987).

The C4.5 program was run 100 times, slowly decreasing the m parameter, which systematically resulted in the learning of deeper and deeper decision trees.[2] Of course, the program would only have had to be run once and analyzed at different points during the decision tree's construction; however, running it multiple times and limiting the depth of learning made the process procedurally easier to study. One could also slowly increase the training set by random sampling, and it has a similar effect as decreasing the m value with a fixed set of training examples. In this case, one could also apply an incremental decision tree learning algorithm (i.e., ID5, see Utgoff, 1988) with a similar control on the parameter m. (C4.5 is not an incremental machine learning algorithm since it does not modify the decision tree on-line with the arrival of new instances. But the decision trees constructed by ID3 and ID5 are identical.) By gradually decreasing m (with a fixed set of training examples), our model assumes that the child's capacity increases in a gradual fashion yielding a series of decision trees in which successors build upon predecessors.[3] Each program invocation roughly corresponds to some fixed period of time, hence, each run will be referred to as an *era* of training.

After each run, the decision tree induced was used to classify the 425 examples which corresponded to the complete set of problems that could be classified into Siegler's six problem types. The responses to 24 problems (four from each of the six problem types) were then used in subsequent analyses to assess the model's success. This is the same set of testing patterns used to evaluate models by McClelland (1989) and Shultz et al. (1995).[4] These patterns reportedly mimicked those of Siegler's original four-peg testing set (McClelland, 1989). It is to these analyses that we now turn.

4.2. Results of the Basic Model

4.2.1. Stage Progression

Figure 2 plots the stage classifications as diagnosed by Siegler's rule assessment methodology for each era of training. For stage 1, this assessment required the correct prediction for balance, weight and conflict-weight test problems and always incorrect predictions for the other problem types. Stage 2 performance required a similar pattern of responses to that of stage 1, except that distance problems also had to be correctly predicted. A stage 3 classification required correct prediction on simple problems and poor performance on conflict problems. A stage 4 classification required that all test problems be solved correctly (Siegler, 1991).

Figure 1 illustrates the pattern of errors predicted by Siegler's rule models. In addition to the predicted distribution of errors, four responses from the test set were allowed to deviate from one of the rules yet still be classified by that rule. This stage classification procedure, including the tolerance for errors, exactly duplicates Siegler's (1976; 1981) methodology.

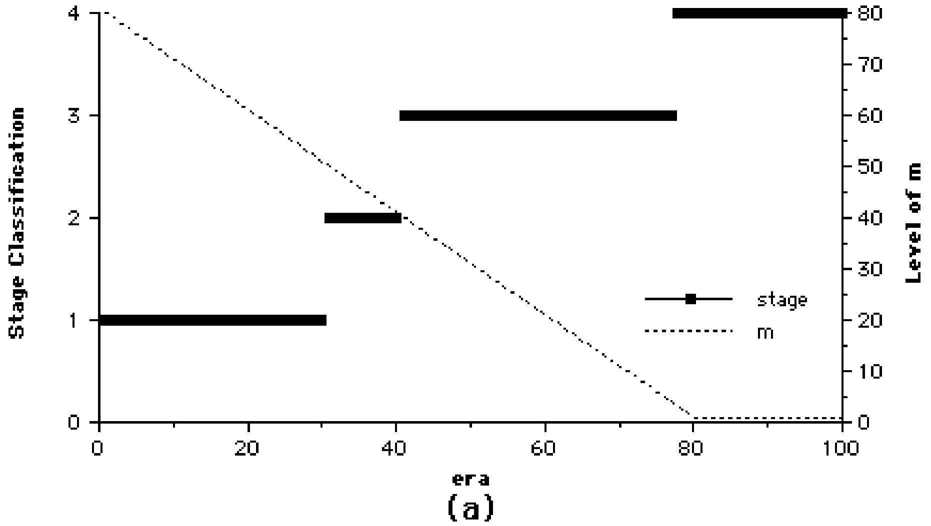
It has been demonstrated that the stage classifications of the rule assessment methodology are not completely orthogonal (Chletsos, De Lisi, Turner, & McGillicuddy-De Lisi, 1989; Shultz et al., 1994). Consequently a set of responses can be classified as either stage 2 or stage 3 depending upon the order in which the assessment methods are applied. Figure 2 (a) plots the decision tree's stage assessment with an attempt to classify each era's responses in the order 4, 3, 2, 1, while Figure 2 (b) plots the assessment in the order 1, 2, 3, 4. Assessment order refers to the sequence of tests used in attempting to classify a set of responses. Success at classifying a set of responses early in the sequence prevents attempts to classify the responses at rule levels occurring later in the sequence. As can be seen from the figures, slightly different rule diagnoses result for certain eras depending upon the order that stage classifications are assessed.

In Figure 2 (a), all four of the developmental stages are represented in succession. In Figure 2 (b), all four stages are present, but with a slight amount of regression. No stage skipping was observed and no eras failed to be classified. Therefore, it is apparent that the C4.5 model has captured the first of the balance scale regularities.

4.2.2. U-Shaped Development

Figure 3 plots the mean longitudinal performance of the simulation on the entire set of conflict-weight problems. The model clearly exhibits U-shaped development on these problems. By comparing the time of occurrence of this performance with the stage classification of the same simulation from Figure 2, it can be seen that the U-shaped developmental trend corresponds precisely with the period in which the simulation is classified at stage 3.

The early reliance on weight information by the simulation is interfered with during this period by the gradual integration and use of distance information on conflict



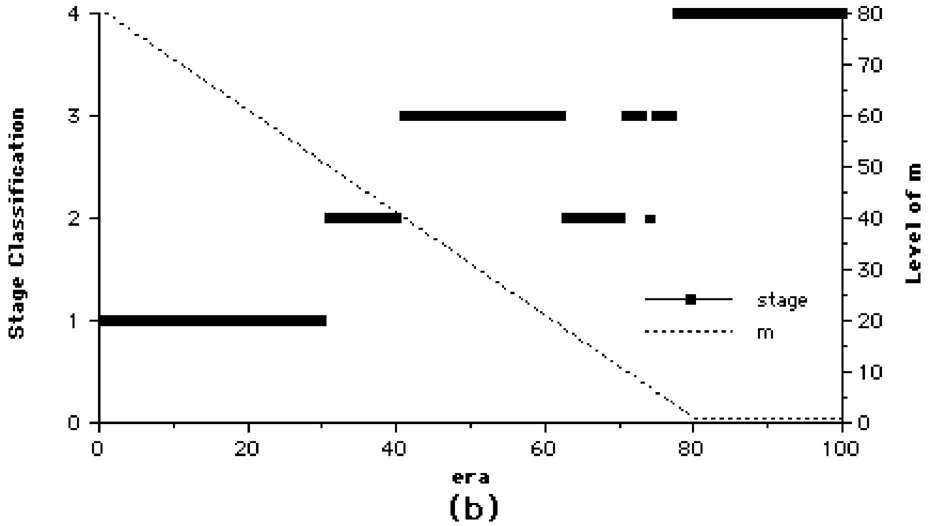


Figure 2. Longitudinal stage progression of Simulation 1 using Siegler's (1976; 1981) rule assessment methodology with stage classification order (a) 4, 3, 2, 1 and (b) 1, 2, 3, 4.

problems. This can be verified by examining the longitudinal performance of the simulation on conflict-distance problems in Figure 4. At precisely the beginning of the period of U-shaped conflict-weight performance, distance information begins to be assimilated. From the diagrams, it appears that there is a gentle vacillation between the learning algorithm's incorporation of weight and distance information

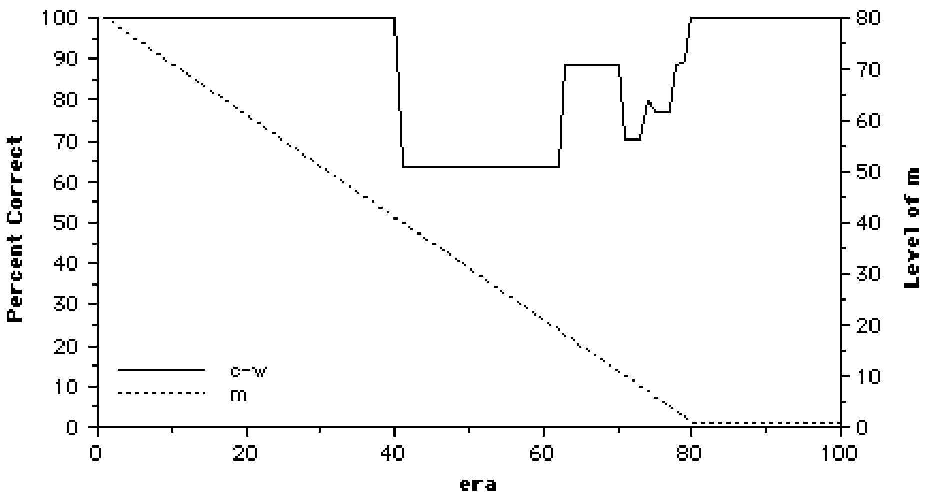


Figure 3. Longitudinal performance on conflict-weight problems.

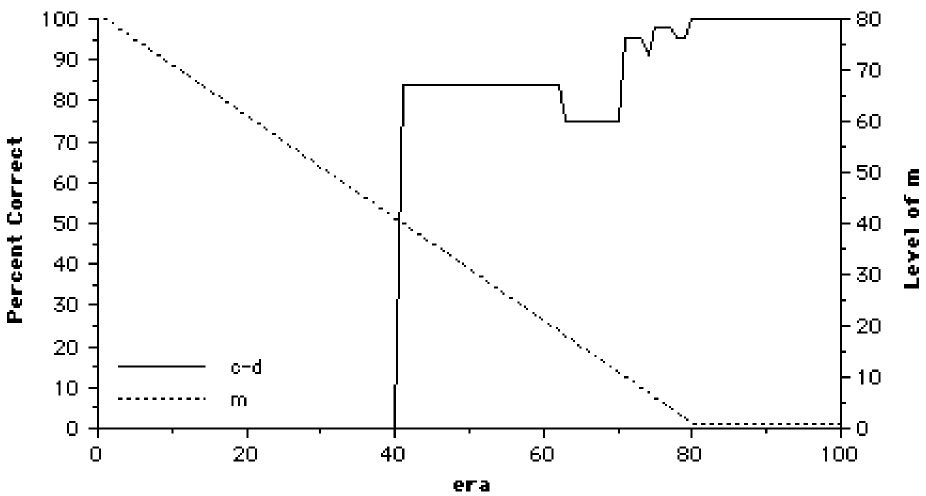


Figure 4. Longitudinal performance on conflict-distance problems.

with the inclusion of information from one of these dimensions conflicting with performance on the other.

4.2.3. Torque Difference Effect

The final major effect characteristic of balance scale development, the TDE, was evaluated in the current simulation by classifying the model's performance using four different sets of testing patterns whose problems were drawn from four different levels of torque difference.[5] The TDE requires that the same set of simulation responses should be classified at different stage levels depending upon the torque difference level of the testing problems. Testing sets with problems from larger torque difference levels should classify the model at a higher stage of development than testing sets with small torque difference problems.

Each testing set had the same balance and conflict-balance testing problems since the torque difference for these types of problems is always zero. The torque difference level for the other testing sets varied. Torque difference level 1 consisted of problems with a torque difference of 1. Levels 2, 3 and 4 consisted of problems with torque differences in the range of 2-5, 6-9 and 10-20 respectively.

We found that only at stage 3, did stage classifications vary in accordance with the predictions of the TDE. This revealed that the simulation was not capturing the TDE at all points in development. Our second simulation will redress this apparent shortcoming of the model.

4.2.4. An Evaluation of the Model's Development

One beneficial aspect of using a symbolic induction method such as C4.5 is that it explicitly represents the knowledge that has been acquired. In this section, we turn to an in depth examination of the decision trees induced throughout the model's ontogeny in order to gain some insight into the model's success, and its inability to capture the full TDE.

Stage 1 in the model was achieved as the result of two distinct sets of rules. The earliest appearing stage 1 rule simply predicted that the side with the greatest weight would go down, and if the weights were equal on either side of the fulcrum, then an outcome of balance was predicted. The second stage 1 rule set first asked whether or not the problem to classify contained equal weights at equal distances. If the problem did conform to this structure then the predicted outcome was balance. Otherwise, the side with the greatest weight was adopted as the response. In the case of distance problems, where weights are equal but distance varies, this rule unrealistically favored one side of the scale.

A single set of rules mapped onto stage 2 performance. These rules built upon the second stage 1 set by considering the side with greater distance when neither side of the scale had greater weight. To give the reader a taste of the style of output that can be achieved from C4.5, this stage's decision tree is reproduced in Figure 5. Readers familiar with the balance scale literature will recognize this decision tree as identical to the stage 2 decision tree first postulated by Siegler (1976) and reproduced in numerous papers on the topic since that time. The first stage 1 rule set was also identical to Siegler's stage 1 rule model, and none of these are capable of demonstrating the TDE.

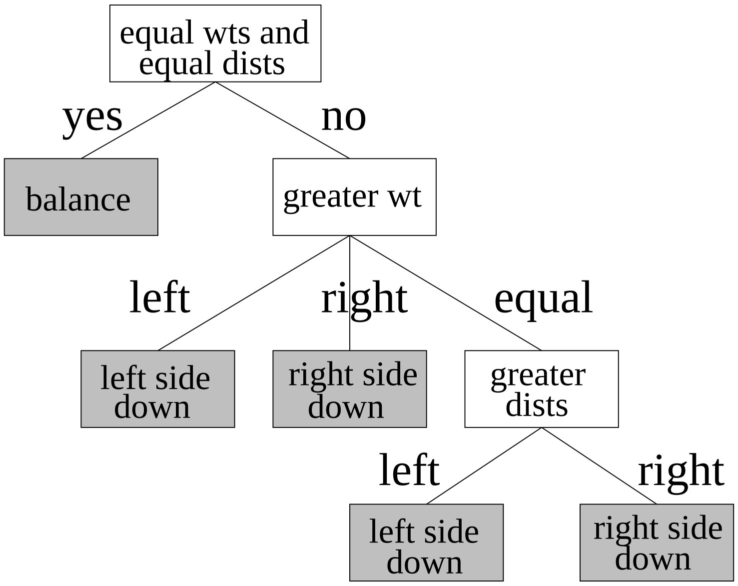


Figure 5. An example of the decision tree output from c4.5. This sample exhibits stage 2 and was produced with $m = 50$. Leaves are shaded.

Stage 3 was accomplished through a set of five distinct rule systems, each built on the previous one. The first stage 3 rule set became active at era 40, and as can be seen from Figures 3 and 4, was immediately responsible for the U-shaped trend in performance on conflict-weight problems and the corresponding consideration of the distance dimension for conflict-distance problems. This decision tree first asked whether weights and distances on either side of the fulcrum were equal, and predicted an outcome of balance if they were. This covered simple balance problems. Next, distance problems were covered by testing if the problem's weights were equal and if so, predicting that the side with the greatest distance would go down. Simple weight problems were covered by predicting that the side with greater weight would go down if distances were equal. In the case of a conflict between sides with greater weight and distance information, if the weight on the side with greater distance was small (less than or equal to 1) then the prediction favored the side with a large weight, otherwise, the side with a larger distance was predicted. This amounts to choosing the side with a larger distance more frequently than the side with the larger weight. Performance on conflict-balance problems is catastrophic since the tree doesn't even consider that weight and distances across the two sides could counteract one another.

The second set of rules corresponding to a stage 3 level of performance is similar to the first, however predictions in the face of conflict problems are further honed. This rule set builds on the previous treatment of conflict problems by considering the magnitude of the distance at which the greater weight is placed. If that distance is small (less than or equal to 2), the prediction favors the opposite side that has its weight (which is at least greater than 1 from the previous test) at a greater distance, otherwise, the side with the greater weight at a distance of at least 2, is favored. This set of rules results in an increased percent correct on conflict problems whose outcomes are on the same side as the greater weight, and this is reflected in Figures 3 and 4.

The remaining three sets of stage 3 rule systems continue to build on this testing of conflict problems which gets more and more particular as C4.5 is allowed to learn to deeper and deeper levels of coverage. From Figures 3 and 4, we can see that the dimensions of weight and distance continue to be integrated, generally with the consideration of one dimension resulting in the loss of correctly responding to problems that rely on the consideration of the other dimension.

The interplay between successively relying more on one dimension than the other continues until stage 4 is achieved. Two distinct decision trees are both classifiable at a level of stage 4 performance. The first does well on all but conflict-balance problems, for which performance is mixed. The second performs flawlessly on all types of balance scale problems, since at this point the parameter $m = 1$, and the entire problem set is covered. No explicit computation and comparison of torques occurs.

4.3. Simulation 2: The Expanded Model

Unless the model discriminates and answers differently, particular problems within Siegler's theoretically equivalent problem types, stage classifications will not vary. Without varying classifications within different problem types, the TDE cannot be observed. If contingencies in the training data exist which distinguish problems based on information other than that used by Siegler (1976; 1981), then the torque difference effect could arise if the learning algorithm were to pick up on such contingencies. Siegler's rule models, and our first simulation's stage 1 and 2 rules all failed to distinguish problems with different input magnitudes. Instead, the induced rules considered only the side of the balance scale with greater weight or distance. Our model's stage 3 rules distinguished problems on the basis of their graded input levels, and its stage classifications did vary with torque difference levels.

It would appear that in order to get the TDE at all stages of development, C4.5 would be required to build rules which discriminated between problems with different levels of inputs. Our examination of the rules induced by simulation 1 revealed that it relied extensively on which side of the balance scale had a greater weight or distance. By representing this feature in an all or none fashion, rules using this information were forced to consider it in an all or none fashion. For our second simulation we augmented our previous model by changing only the representational format of which side of the balance scale had greater weight or distance. In this second simulation these variables were treated as continuous variables[6] and they took on values in the range of $-4 \leq x \leq 4$ (determined by subtracting the right side value from the left side value for each of the weight and distance dimensions). By doing this, we have prevented C4.5 from being able to consider the side of the balance scale with larger weight or distance information in an all or none fashion, and instead have forced it to consider the attribute in terms of a graded representation. No other conditions of the model were altered, and training and assessment were carried out as in simulation 1.

4.4. Results of the Expanded Model

The second simulation was first examined for the TDE by assessing each era independently with four different sets of testing problems drawn from the four different torque difference intervals outlined earlier. Stage classifications varied on each of the first 77 eras demonstrating that the simulation exhibited the TDE throughout its earliest period of development. Such an effect is also present in the human data. Beyond era 77, the simulation reached a saturation point for the training problems, and all of the problem sets were classified at a stage 4 level of performance.

The simulation's overall performance was also evaluated on the entire set of problems in the four torque difference ranges. This was done by calculating the percentage of correct responses at the median era of each stage. This amounted to evaluating the model at eras 3, 25, 43 and 79 for stages 1, 2, 3 and 4 respectively. As dictated by the TDE, the model demonstrated superior performance on problems from larger torque difference intervals. The results of this assessment appear in Figure 6. From the results of these analyses for the TDE, it is clear that the model is sensitive to torque difference.

An examination of other aspects of the model's performance revealed that as in Simulation 1, every stage was classifiable, and the simulation demonstrated orderly stage progression. Longitudinal conflict-weight performance showed the characteristic U-shaped regression in performance that coincides with stage 3, however conflict-weight performance at the very earliest stage of development was slightly poorer. Nonetheless, the expanded model captured all of the major aspects of the developmental data.

Rules induced in the second simulation were more complex because they included tests based on a larger number of gradations in the attribute set. Figure 7 shows an example of the rule induced in the expanded model when $m = 50$. This can be contrasted with the rule presented in Figure 5 from the basic model. The expanded model rule first considers whether weights and distance are equal for the problem at hand, and if so, predicts that the scale will balance. Otherwise, the gradation level of difference between the arms for the weight dimension is considered. Depending upon the outcome of considering the weight dimension, the distance dimension is tested for both the arm on the same and the opposite side of the scale, and a response is derived.

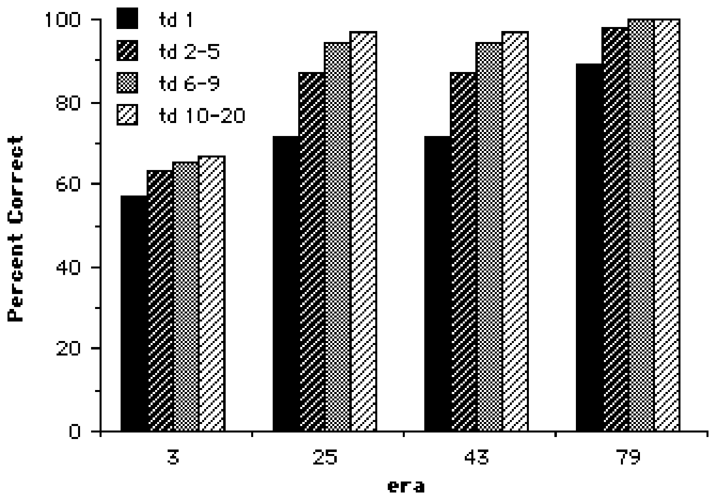


Figure 6. Percent correct on all problems at four torque difference levels sampled at the median epoch of each stage classification for Simulation 2.

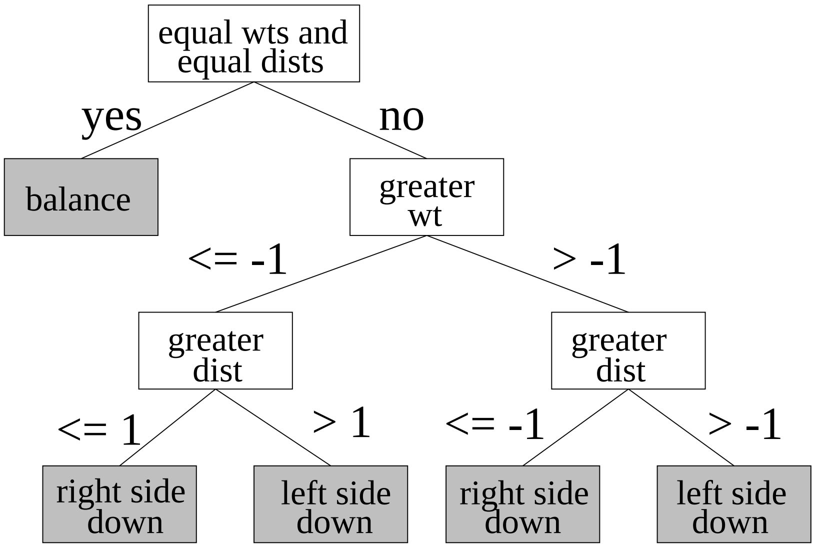


Figure 7. An example of the decision tree output from the expanded model. This sample exhibits stage 3 performance and was produced with $m = 50$. Leaves are shaded.

From this example rule, it is clear that from early on, the rules induced are making use of the graded information that has been supplied; yet the resulting tree's performance is still consistent with Siegler's four rule models.

Table 1. Summary of published balance scale models to date.

Model	Stages	U-Shaped Development	TD Effect	Stage Skip | Regression	Scalable Model	Number of Training Instance	Assessment Methods	Assumptions
Klahr & Siegler	none	n/a	n/a	n/a | n/a	n/a	n/a	n/a	n/a
Langley (discrimination learning)	maybe	no	no	n/a | n/a	no	110	qualitative percentage correct	supervision
Newell (SOAR)	1,2,3	yes	no	probably | doubtful	no	5	qualitative based on 5 problems	supervision
McClelland (BP)	1,2,3,4	yes	no	yes | no	no	10000	rule assessment, torque difference	supervision, variable training set, bias for weight dimension, hidden unit segregation, interpreted outputs
Shultz & Schmidt (CC)	1,2,3,4	yes	yes	yes | yes	maybe	11000	rule assessment torque difference	supervision, gradually expanding training set, bias for weight dimension, interpreted outputs
Shultz, Schmidt, Buckingham, & Mareschal (CC)	1,2,3,4	yes	yes	yes | yes	no	62500	rule assessment, torque difference	supervision, innate bias, interpreted outputs
Schmidt & Ling (C4.5)	1,2,3,4	yes	yes	no | some	yes	675	rule assessment, torque difference	supervision, bias for balance problems, capacity increases with age

5. General Discussion

Previous symbolic accounts of the balance scale task have been lacking a transition mechanism capable of delivering the major phenomena of regular stage progression, U-shaped development on conflict-weight problems and the TDE. Connectionist models on the other hand, have been successful at capturing all of these effects.

Our symbolic C4.5 model assumed that balance problems are especially salient to children, and that the majority of children are internally biased towards processing the weight dimension over the distance dimension. In addition we assumed that children have access to information about which side of the balance scale is larger for a given dimension, and that they use this information during reasoning. By implementing these assumptions and applying the C4.5 learning algorithm, the model presented in the current paper provides an alternative developmental model, capable of successfully capturing many aspects of the human data.

Table 1 presents a summary of the major balance scale models to date, along with a concise listing of their capabilities and assumptions. Because the majority of the information found in this table has been discussed throughout the current paper, the table will not be referenced further.

5.1. Robustness of Assumptions Underlying Symbolic and Connectionist Solutions

The C4.5 algorithm and attribute set presented provides a rather robust set of assumptions for modeling variations of the current balance scale task. McClelland (1989) generated a model of a five-peg, five-weight balance scale, and the other connectionist researchers followed suit. This was despite the fact that Siegler's (1976; 1981) original balance scale data was based on a four-peg, four-weight version of the problem. Pilot experiments using four-peg, four-weight and six-peg, six-weight versions of McClelland's model with back-propagation failed to achieve a satisfactory result. Similar results were observed with pilot cascade-correlation networks. In both cases only stage 3 and 4 behavior occurred. McClelland (personal communication) suggested that in the case of the four-peg, four-weight version of his model, the training set was too small for the model to learn in a realistic fashion. In modeling the larger-sized balance scale, the networks differentiate the more fine-grained inputs in a manner that does not yield a psychologically realistic simulation.

Our C4.5 model on the other hand, works just as well for smaller and larger balance scales as it does for the five-peg, five-weight version. We created a four-peg, four-weight basic model of the balance scale task. The first four sets of decision trees were exactly the same as those reported in simulation 1. Two different decision trees were responsible for stage 1 performance, one for stage 2, three for stage 3, and two for stage 4. Fewer trees were required because stage 3 took less effort to capture. Late in stage 3, rule sets differing from the larger five-peg, five-weight balance scale model appeared. These rules differed in the magnitudes of the tests used on the weight and distance values, as well as the order in which the attribute information was examined.

A six-peg, six-weight version of the model was also successful. This model induced one decision tree corresponding to stage 1 and a second for stage 2. Four decision trees resulted in stage 3 performance and two for stage 4. At least the first three trees induced were the same as those for the five-peg, five-weight model. Later rules differed in terms of the order in which they used weight and distance information and the magnitudes that they differentiated problems by. It is still an open empirical question whether balance scale data of other sizes can be accommodated with connectionist techniques.

C4.5 is also robust with respect to the format of its output encoding. While connectionist models' success hinge on the inclusion of a distributed encoding of two outputs (an architectural assumption, see McClelland, 1989), with the C4.5 model, alternative methods of representing the response yield identical results.

The fact that existing connectionist models fail to accommodate balance scale data from scales of other sizes, and crucially rely on a specific output format, suggests that their success at modeling the five-peg, five-weight version of the balance scale task may be a serendipitous by-product of their construction, as opposed to the result of crucial properties supplied by the learning algorithms or the models' assumptions.

We attempted to train Cascade-Correlation with our attribute and training set. Only stages 3 and 4 ensued. Encoding the outputs in a distributed fashion as required by previous connectionist models also yielded only stage 3 and 4 behavior. Stages 1 and 2 were not observed because there was no bias in the training set towards weight information. This finding, coupled with our earlier report of a failure to observe a successful model using C4.5 and the connectionist training inputs, underscores the notion that representational and learning biases of different machine learning algorithms vary, and that the algorithm used is a crucial component of any model.

Finally, in contrast to the vast space of possible connectionist implementations which possess an enormous number of degrees of freedom and require the tweaking of a large number of free parameters, the C4.5 model varied only a single free parameter, m. While connectionist implementations have the capability of closely matching the human performance data, the degree to which their solutions provide a match is dependent upon simultaneously setting a large number of free parameters (Schmidt & Shultz, 1992). With only a single parameter to tweak, there are more restrictions when using C4.5.

5.2. Competence versus Performance

There are a number of differences between the modeling approach taken in the current paper and that of previous connectionist accounts. The latter models attempt to duplicate human performance in a very fine grained manner. Our account is more concerned with characterizing the knowledge structures underlying human performance. Chomsky (1968) differentiated models of *competence* from models of *performance*. Competence is the ability of an idealized subject to execute the task at hand. This ideal is not affected by situational variables, memory span, or perceptual limitations. In reality, competence is revealed only indirectly through a subject's performance, which is always influenced by situational factors. As a result, a subject's actual performance seldom equates to their competence.

The C4.5 model, as presented in the current paper is intended as a competence model. It learns from the entire set of balance scale problems, and it always follows the same course of development. Similarly, unlike the human situation, the model is unconstrained in the devotion of its resources to the problem at hand. Clearly, these are idealized learning and performance conditions.

If the model is examined as a model of balance scale performance, there are a number of disturbing features of the C4.5 account. First, changes are rather abrupt since groups of problems are solved with each new addition to the maturing decision tree. Nonetheless, it is still the case that multiple sets of decision trees map onto a single stage of performance (i.e., stage 1 behavior results from two separate decision trees, stage 3 behavior from five, and stage 4 behavior from two) demonstrating that development is incremental. Since there are no undetectable fine grained incremental changes in learning, the model seems not to be supported by reports that balance scale learning can occur on a problem by problem basis (Kliman, 1987; McClelland & Jenkins, 1991).

A second aspect of the performance data that the C4.5 model does not capture concerns the issue of variability in the model's output. Stage skipping, regression, lack of subject classification, and individual differences all mark the human developmental data. Our model did not exhibit any eras which were unclassifiable by Siegler's rule diagnosis methods. It also did not exhibit any stage skipping or regression when assessed with the rule order 4, 3, 2, 1. Furthermore, the model failed to be able to accommodate individual differences — each run is the same.

Measures could be taken to augment the workings of C4.5 in an effort to reduce its impeccable performance in an attempt to make it accountable for human performance as well as competence. For instance, a promising first attempt would be to modify C4.5 to build its decision trees by sampling probabilistically from the training set. In this scheme, the edges disseminating from each node would have a probability associated with them that reflects the likelihood that each would lead to what is assumed to be the correct response for a given developmental level. The probability levels are induced along with the tree, and reflect the structure of the learning domain. The added variability would therefore be based on the data and not simply result from random variability.

A second method of introducing variability into the current model would be to restrict the number of training instances that the learning algorithm operates on. Pilot experiments with the C4.5 model have shown that by training on a randomly selected group of exemplars each era, the C4.5 model can produce a failure to classify all of the eras and also to witness stage regressions and skipping. Figure 8 shows a plot of the longitudinal stage classifications for a simulation which randomly selected 500 training instances without replacement, during each era of training. This manipulation obviously degrades the perfect performance of the unaffected model. One might also pursue a performance model using an incremental decision tree learning algorithm (i.e., ID5, see Utgoff, 1988) with a similar control parameter such as m and injecting variability through the use of randomly chosen training exemplars.

Finally, it should be noted that the use of explicit representations in a competence model does not imply that symbolic modeling methods are only appropriate for tasks in which subjects have conscious access to explicitly represented knowledge. Simply because propositions are assumed to take part in the execution of a task does not necessitate that subjects have conscious access to the information that is represented within those propositions (Chomsky, 1968). Implicit tasks (of which it may be hypothesized that the balance scale task is an instance, see McClelland, 1995) can also be modeled using symbolic methods. Ling and Marinov (1994) have successfully applied C4.5 to a pair of famous implicit learning tasks, and one might consider the current balance scale model as yet another such application.

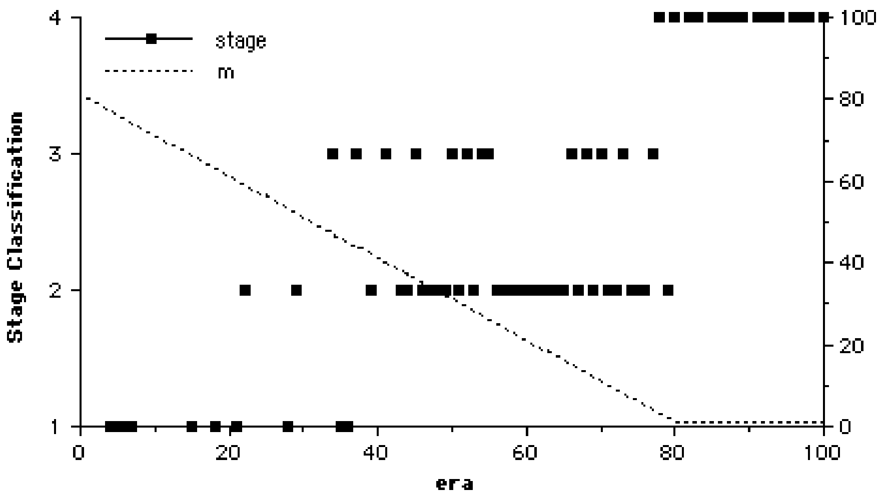


Figure 8. Longitudinal stage progression of a pilot simulation using Siegler's (1976; 1981) rule assessment methodology with stage classification order 1, 2, 3, 4. This pilot simulation was trained on 500 randomly chosen training instances and clearly exhibits stage skipping, regression and unclassified eras in training.

5.3. Implications for the Torque Difference Effect

An intriguing finding with C4.5 was its ability to capture the TDE from within a rule-based system. Previous symbolic modeling efforts seem unlikely to have demonstrated this phenomenon due to their failure to incorporate sufficiently graded representations. The human literature hypothesized that the TDE was a result of information salience: Ferretti and Butterfield (1986) suggested that when distance differences between balance scale arms were large on a given problem, children were compelled to integrate information from that dimension.

Shultz et al. (1994) proposed that connectionist models show the TDE because they are naturally sensitive to weight and distance differences. It was also argued that rule-based models would fail to show the TDE if they were insensitive to these amounts. Continuous computation was deemed necessary to observe the TDE. Sensitivity, in this context, presumably refers to making decisions based on numerical magnitudes. Initially, it would appear that the C4.5 model was doing just that — we gave it information about the problems in terms of the difference in input magnitudes on each of the weight and distance dimensions, and the TDE obtained. However, it should be noted that whether C4.5 treated this information as a continuous variable or as a discrete symbol, the same result occurred. For C4.5, it was not the magnitude of the difference responsible; in fact the learning algorithm itself is insensitive to these amounts. Rather, the graded composition of the training examples was responsible for this effect.

An examination of the torque difference levels in the training corpus shows that there are more problems of smaller torque differences than there are of larger torque differences. This suggests that larger torque difference problems may be easier for the learning algorithms to discriminate. Based on our C4.5 results and the fact that C4.5 reduces variability in the training data based on an information gain measure, we can conclude that there was greater information gain by covering larger torque difference problems before covering smaller problems. As a result, performance on these problems was better at all times in development.

Our results suggest that the TDE is a byproduct of learning methods which acquire new information by maximizing error reduction coupled with a sufficiently fine grained representation of the training data. We know that C4.5 uses such a criterion to decide when and how to expand its existing decision tree. Similarly, connectionist learning algorithms learn such that they are continually reducing some global measure of error. These methods are guaranteed to follow the path of steepest descent on a local error surface, which is essentially maximizing their ability to account for an increasing number of the training examples. Perhaps it is this feature of both connectionist methods and the C4.5 learning system, coupled with the structure of the problem domain, that is responsible for their ability to generate accurate balance scale models featuring the TDE. Such an induction technique may also characterize the human learner.

5.4. Predictions

The C4.5 model makes a number of predictions that are different than or opposed to those made by previous accounts. First, while many connectionist accounts assume an environment strongly biased towards presenting information about the weight dimension (McClelland, 1989; Shultz et al., 1994), the C4.5 model predicts that the weight and distance dimensions are equally and symmetrically presented in the natural world.

Additionally, the C4.5 model adopts an internal preference for either weight or distance information. If this characterization is correct, then some human subjects might naturally prefer the distance dimension over the weight dimension, opposite to the commonly assumed trend. To account for such data, the aforementioned connectionist models would find themselves in the contradictory position of assuming an environment, shared by all subjects, that is biased for both weight and distance information. To escape the contradiction, one might suppose that the source of the bias is internal to the subject, as we have, but doing so alters the current claims that the source of the bias is environmental (McClelland, 1989; Shultz et al., 1994). Alternatively, one might assume that the source of the bias is environmental, and that each child has an environment uniquely biased towards either weight or distance information (Shultz, personal communication). While this is an alternative view, adopting such a position requires an account of the sources of such information, and why environmental differences might exist.

An empirical test as to whether all children share an environment biased towards gaining experience with weight information might consist of examining the balance scale responses of a large human population for evidence of treating either the weight or distance dimensions as principle. If variability exists in the principle dimension, then a single environmental source of the bias is highly unlikely. A preliminary report by Chletsos, et al. (1989) suggests that such variability can exist, as it does in other related tasks involving the integration of information from two dimensions (i.e., inclined plane, projection of shadows, conservation, class inclusion, fullness, and several other tasks, see Siegler, 1976; 1981; 1991). It should be noted that models assuming an environmental bias cannot be applied directly to such tasks where variability in the principal dimension is known to exist.

A second prediction of the C4.5 model is that problems in which equal weights occur at equal distances from the fulcrum thereby resulting in balance outcomes, are particularly salient. Johnson (1987) has discussed extensively the important and fundamental role that balancing plays in cognitive development. An interesting line of empirical research could experimentally examine some of Johnson's claims in order to discover whether balancing is as primary as his work and our model supposes.

A third prediction of the C4.5 model is that children reason with primitive information about the balance scale problem that they are trying to solve. Part of the input to the learning algorithm indicated whether the problem had equal weights at equal distances, and which side of the balance scale dominated each of the dimensions of weight and distance. Primitive information of this sort, which can easily be determined from the visual presentation of a balance scale problem, was used by Siegler's (1976; 1981) decision trees models of children's reasoning, and also played an important role in the decision trees that C4.5 induced. The model therefore predicts that such primitive information can play an important role in children's reasoning. Preliminary empirical work on this topic suggests that children do reason about balance scale problems using primitives such as these, as well as others (Kliman, 1987). It should be stressed however, that unlike children, the current model does not induce these primitives on its own.

A fourth prediction that the model makes about human cognitive development concerns the way in which information is integrated from two dimensions longitudinally. Siegler's stage 3 decision tree has been criticized on a number of accounts. For conflict problems, Siegler suggested that children "muddle through" or respond randomly at this stage. Other researchers attempted to examine stage 3 performance for more systematic behavior (Ferretti, Butterfield, Cahn, & Kerkman, 1985; Normandeau, Larivee, Roulin, & Longeot, 1989; Wilkening & Anderson, 1982). This empirical research found that children's stage 3 behavior could be classified by a number of different rules and, therefore, behavior was more systematic than had been assumed.

The C4.5 model presents a number of differing rules capable of yielding stage 3 classifications. During stage 3, children regress on conflict-weight problems. Our earlier examination of the rules responsible for this regression revealed that it was due to the preliminary integration of distance information of small magnitudes. This step was to the detriment of behavior on conflict-weight problems. Later stage 3 decision trees alternately relied more on weight and then distance information. That is, subsequent rules whose performance was classified at stage 3 vacillated between fine tuning the decision tree's performance by introducing new decisions based on either weight or distance information.

A comparison of Figures 3 and 4 reveals that the longitudinal effect that changes in the decision tree had was to increase performance on the newly introduced dimension at the expense of performance on the opposite dimension. Therefore, the model predicts that stage 3 performance is systematic, and that it will vacillate longitudinally between performing well on weight and distance conflict problems. Furthermore, the model predicts that during stage 3, there will be a number of regressions in performance on conflict problems, separated by periods of improved performance, finally culminating in getting all of the conflict problems correct by stage 4.

A final prediction of the C4.5 model is that the TDE may not be the result of perceptual salience as previous empirical and theoretical explorations have suggested, but the result of a graded representational format coupled with a greedy learning mechanism (be it connectionist or symbolic). Hence, the structure of the problem domain, coupled with the style of learning, is hypothesized to be responsible for observing the TDE. This account does not rule out that perceptual salience may contribute to, or exaggerate, the effect in humans. This prediction is as much for modelers attempting to capture the TDE as it is for cognitive development researchers.

6. Conclusion

In summary, we have presented a model of human cognitive development on the balance scale task using the symbolic C4.5 learning algorithm to induce decision trees capable of capturing the major trends present in the human developmental data. This model stands in contrast to the previously limited symbolic approaches to the problem, providing a successful alternative set of assumptions and competing predictions, to connectionist accounts of the data. Our demonstration suggests that symbolic learning algorithms warrant investigation as transition mechanisms in cognitive development.

Acknowledgments

This research was supported by NSERC Canada Post Graduate funding and an Honorary Izaak Walton Killam Scholarship to William C. Schmidt, as well as a

NSERC Canada Research Grant (OGP0046392) to Charles Ling. The manuscript benefited greatly from the comments of Michael Pazzani, Jeff Shrager, Thomas Shultz, Lisa Tellier and two anonymous reviewers.

Notes

1. The six types of problems do not include problems with greater weight and greater distance on the same side.
2. We began with $m = 80$ and decremented m with each running of C4.5 until it reached the value 1. For the sake of aesthetics, and to reify that stage 4 was stable, we continued to assess the model with $m = 1$ for another 20 runs.
3. Note that when m decreases, a node that was not considered for splitting with larger values of m may have larger information gain than nodes considered for splitting earlier. In this case, the decision tree constructed may not expand on the earlier ones. In our simulations, this happens only once; the initial tree of simulation 1 was not a precursor to the rest. (Section 4.2.4).
4. The testing set is reproduced in Schmidt & Shultz (1991). Thanks to Jay McClelland for making his testing set available.
5. The specific testing problems used are reported in Schmidt & Shultz (1991).
6. The use of a discrete, symbolic representation yields identical results, however this prevents generalization to continuously represented problems that the model was not trained on.

References

Chletsos, P. N., De Lisi, R., Turner, G., & McGillicuddy-De Lisi, A. V. (1989). Cognitive assessment of proportional reasoning strategies. *Journal of Research and Development in Education*, **22**, 18–27.

Chomsky, N. (1968). *Language and Mind.* New York: Harcourt, Bruce and World.

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

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. *Child Development*, **39**, 131–160.

Hunt, E., Marin, J., & Stone, P. (1966). *Experiments in Induction.* New York: Academic Press.

Johnson, M. (1987). *The Body in the Mind: The Bodily Basis of Meaning, Imagination, and Reason.* Chicago: The University of Chicago Press.

Klahr, D., & Siegler, R. S. (1978). The representation of children's knowledge. In H. W. Reese & L. P. Lipsitt (Eds.), *Advances in child development and behavior* (pp. 61–116). New York: Academic Press.

Kliman, M. (1987). Children's learning about the balance scale. *Instructional Science*, **15**, 307–340.

Langley, P. (1987). A general theory of discrimination learning. In D. Klahr, P. Langley, & R. Neches (Eds.), *Production system models of learning and development*, pp. 99–161. Cambridge, MA: MIT Press.

Ling, C. X., & Marinov, M. (1993). Answering the connectionist challenge: a symbolic model of learning the past tenses of English verbs. *Cognition*, **49**, 235–290.

Ling, C. X., & Marinov, M. (1994). A symbolic model of the nonconscious acquisition of information. *Cognitive Science*, **18**, 595–621.

McClelland, J. L. (1989). Parallel distributed processing: Implications for cognition and development. In R. Morris (Ed.), *Parallel Distributed Processing: Implications for Psychology and Neurobiology.* Oxford: Clarendon 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.* Hillsdale, NJ: Erlbaum.

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.* Hillsdale, NJ: Erlbaum.

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

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

Quinlan, J. R., (1986). Induction of decision trees. *Machine Learning,* **1**, 81–106.

Quinlan, J. R., (1993). *C4.5: Programs for machine learning.* San Mateo, California: Morgan Kaufmann.

Sage, S., & Langley, P. (1983). Modeling cognitive development on the balance scale task. *Proceedings of the Eighth International Joint Conference on Artificial Intelligence,* **1**, 94–96. Karlsruhe, West Germany.

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

Schmidt, W. C., & Shultz, T. R. (1992). An examination of balance scale success. In *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society,* pp. 72–77. Hillsdale, NJ: Lawrence Erlbaum.

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

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

Shultz, T. R., Schmidt, W. C., Buckingham, D., & Mareschal, D. (1995). Modeling cognitive development with a generative connectionist algorithm. In T. Simon & G. Halford (Eds.) *Developing cognitive competence: New approaches to process modeling.* Hillsdale, NJ: Erlbaum.

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

Siegler, R. S. (1981). Developmental sequences between and within concepts. *Monographs of the Society for Research in Child Development,* **46** (Whole No. 189).

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

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

Utgoff, P. E. (1988). ID5: An incremental ID3. In *Proceedings of the Fifth International Conference on Machine Learning.* Morgan Kaufmann.