University of Alberta

A Network Interpretation Approach to the Balance Scale Task

by

Corinne Lynn Zimmerman ©

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of *Doctor of Philosophy*

Department of Psychology

Edmonton, Alberta
Fall 1999

0-612-46955-7

University of Alberta

Library Release Form

Name of Author: Corinne Lynn Zimmerman

Title of Thesis: A Network Interpretation Approach to the Balance Scale Task

Degree: Doctor of Philosophy

Year this Degree Granted: 1999

Corinne Zimmerman

Box 695
Melville, Sask.
S0A 2P0

31 August 1999

University of Alberta

Faculty of Graduate Studies and Research

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research for acceptance, a thesis entitled *A Network Interpretation Approach to the Balance Scale Task* submitted by *Corinne Lynn Zimmerman* in partial fulfillment of the requirements for the degree of *Doctor of Philosophy.*

Michael R. W. Dawson
Co-supervisor

Gay L. Bis

Gay L. Bisanz
Co-supervisor

Connie Varnhagen

Walter Bischof

Michael Carbonaro

Michael Carbonaro

Thomas R. Shultz
External Examiner

AUG.31/99

Abstract

The focus of this research is the balance-scale task of naive or intuitive physics first introduced by Piaget. Developmental psychologists have been interested in this task because of age-related trends in performance, a U-shaped trend on a particular class of problems, and information salience effects. The balance-scale task has also become a benchmark task for connectionist researchers interested in modeling cognitive development. The standard approach to modeling has involved manipulating various parameters and outlining particular assumptions in order to get network models to display behavior that matches the psychological data. A novel approach to studying the balance-scale task is presented. Neural networks were trained to make balance scale predictions and then the converged networks were interpreted using a number of techniques. This approach was taken to gain insight into the nature of how humans solve the task by first examining how neural networks solve the task and by examining the characteristics of the task. Predictions were derived from the analysis of the neural networks and the problem space analysis. In particular, predictions focused on the idea that performance measures (accuracy and RT) should vary as a function of where a problem is located in the problem space. This prediction was tested with a group of undergraduates who were aware of the importance of both the weight and the distance dimension for making predictions, but who were not familiar with the mathematically correct rule. Intra-individual variability was found in RT and accuracy measures. This contrasts with typical descriptions of balance scale performance as consistent with a single strategy per developmental stage. Implications for future research are discussed.

Acknowledgement

I would like to thank the members of my supervisory committee: Michael Dawson, Gay Bisanz, and Connie Varnhagen for their constant support and encouragement. I would also like to thank Walter Bischof for serving as chair of the examining committee and for informal mentorship over the years.

I would like to thank Jacqueline Leighton, Jan Snyder, Monica Valsangkar-Smyth and Leanne Willson for glee club meetings (both shop and non-shop talk). Special thanks to Billy Schmidt for invaluable conversations about my research, to David Medler for inviting me to the BCP, and to David McCaughan for teaching me how to manoeuver around the BCP. I would like to thank Rob Vogt and Barry Posner for putting up with me when I wasn't in the lab. Leslie Twilley deserves special acknowledgement for her help and support from Day 1.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

B, BAL	Balance.
BG	Both-Greater problems.
CB	Conflict-Balance problems.
CD	Conflict-Distance problems.
CW	Conflict-Weight problems.
CBSD	Conflict-Balance Sum-Distance (torque rule predicts balance, additive rule predicts side with greater distance will tip).
CBSW	Conflict-Balance Sum-Weight (torque rule predicts balance, additive rule predicts side with greater weight tips).
CDSB	Conflict-Distance Sum-Balance (torque rule predicts side with greater distance tips, additive rule predicts balance).
CDSW	Conflict-Distance Sum-Weight (torque rule predicts side with greater distance tips, additive rule predicts side with greater weight tips).
CWSB	Conflict-Weight Sum-Balance (torque rule predicts side with greater weight tips, additive rule predicts balance).
CWSD	Conflict-Weight Sum-Distance (torque rule predicts side with greater weight tips, additive rule predicts side with greater distance will tip).
D, DIS	simple Distance problems.
L	Left.
LD	Left Distance.
LW	Left Weight.
R	Right.
RD	Right Distance.
RT	Reaction Time.
RW	Right Weight.
TD	Torque Difference ($\lvert RW \times RD - LW \times LD \rvert$) .
TDE	Torque Difference Effect.
W, WT	simple Weight problems.

CHAPTER 1

INTRODUCTION

The balance-scale task of intuitive physics has been of interest to developmental psychologists since its introduction by Piaget (e.g., Inhelder & Piaget, 1958). The task involves making a prediction about the state of a two-armed balance (i.e., tip or balance) based on a configuration of weights at particular distances from the fulcrum. The task is appealing because of age-related trends in performance, a U-shaped trend on a particular class of problems, and information salience effects. As the intersection between research on cognition, development, and connectionism grows, the balance-scale task has emerged as a benchmark problem for researchers attempting to model cognitive development (Shultz, Mareschal, & Schmidt, 1994; Shultz, Schmidt, Buckingham, & Mareschal, 1995).

The balance-scale task, like many tasks that have been used to study cognitive development, has two key characteristics (Siegler, 1996). First, the task as administered is not particularly familiar to children. Children have experienced the concept of balance but typically have not encountered this type of prediction task. The rationale is that by using novel problems, we can learn about participants' naive conceptions and the strategies they employ when faced with unfamiliar problems. Second, it is among a class of tasks that involves the integration of information from two dimensions. Other examples include conservation of number, conservation of liquid, the slopes task, and the projection of shadows task (Siegler, 1976; Wilkening & Anderson, 1982).

The central thesis of this dissertation is that the balance-scale task is subject to the same inter- and intra-individual variability that is characteristic of the more *familiar*, everyday tasks studied by cognitive developmentalists such as math (e.g., Bisanz & LeFevre, 1990; Siegler & Crowley, 1991), reading (e.g., Perfetti, 1992), spelling (e.g., Varnhagen, 1995; Varnhagen, McCallum, & Burstow), time telling (Siegler & McGilly), and scientific reasoning (e.g., Kuhn, Garcia-Mila, Zohar, & Andersen, 1995; Schauble, 1990, 1996; Schauble & Glaser, 1990). The developmental course for these tasks has been described with the "overlapping waves" depiction of development (e.g., Siegler, 1995, 1996). At any one time, an individual has a variety of available strategies. Rather than sudden shifts from one qualitatively different way of thinking to another, change occurs through competition among strategies.

The balance-scale task, in contrast, has been characterized according to an older view of development: each stage of development corresponds to a single problem-solving strategy. Siegler (1996) has appealed to the unfamiliarity of the balance-scale task as a reason for why it is that a single, consistent strategy is observed for individuals at different stages of development. As will be shown from the review of literature in Chapter 2, however, many factors have been shown to affect the evaluation of individuals as using a single consistent strategy. In fact, it was the numerous criticisms of the rule-assessment method that prompted the research in this dissertation. Because rule assessments for an individual could vary with a number of different factors (e.g., task demands, the particular items used in the testing set), the method has been subject to criticism. Despite the growing number of criticisms, modeling researchers using either production system or connectionist architectures have attempted to produce models that provide a good fit to the human data collected with the rule-assessment method. Moreover, modelers have used the rule-assessment method to evaluate the performance of their models.

In summary, the main argument advanced in Chapter 2 is that the numerous criticisms of the rule-assessment method have made the human data open to alternative interpretations. As such, computer models designed to capture the regularities in this data are also open to interpretation. Therefore, a new approach is advocated--an approach based on interpreting neural networks. This approach was taken to gain insight into the nature of how humans solve the task by first examining how neural networks solve the task and by examining the characteristics of the task.

In Chapter 3, the results of the network interpretation approach will be presented. Neural networks were trained to make balance-scale predictions and then the converged networks were subjected to four main interpretive techniques. A key finding was that the network was integrating the weight and distance dimensions and solving the task by approximating an additive function (although the mathematically correct method requires multiplication). An analysis of the problem space revealed that the majority of balance-scale problems could, in fact, be solved using an additive heuristic. This finding motivated an analysis of the characteristics of previously published test sets and an additional simulation in which neural networks were trained without the problems that cannot be solved using an additive heuristic.

In Chapter 4, predictions were derived from the analysis of the neural networks and the problem space analysis. In particular, predictions focused on the idea that performance measures (accuracy and RT) should vary as a function of where a problem is located in the problem space. The neural networks in Chapter 3 responded differentially to problems depending on location in the problem space. This prediction was tested with a group of undergraduates who were aware of the importance of both the weight and the distance dimension for making predictions, but who were not familiar with the mathematically correct rule (i.e., the *torque rule*, which involves a comparison of the product on the left side [weight x distance] with the product on the right side). Torque difference was used as one rough index of location in the problem space (i.e., the absolute value of the difference in torque on the left and right sides). Torque difference was a good predictor of both accuracy and reaction time measures. Differences in accuracy and reaction time were found on the subset of items that can only be solved via the torque rule.

As mentioned, the criticisms of rule-assessment motivated a novel approach to studying the balance-scale task with respect to both neural networks and human performance. During the course of this research, it became clear that the criticisms of rule assessment should not be focused on it being an inadequate method of evaluation, but rather they should be focused on the *goals* of the rule-assessment approach. In Chapter 5, I argue that the goal of rule-assessment has been to determine the one consistent or modal strategy for individuals at different stages of development. Siegler (1996) has appealed to the *moderate experience hypothesis*, citing the unfamiliarity of the task as the main reason why variability is not seen. That is, he has provided an explanation for why variability in strategy is not observed on the balance-scale task, when in reality, the cumulative evidence points to variability in strategy based on characteristics of the particular instance (e.g., the torque difference). This claim must be qualified, however, as the present results bear only on individuals who are aware that the two dimensions of the task must be integrated. The claim in general is supported by previous research in the literature in which variability in assessment was found with children of different ages.

Chapter 5 ends with a discussion of implications of the present research, suggestions for extensions to analogous developmental studies, and speculations about how current conceptions of development will influence future modeling studies.

CHAPTER 2

LITERATURE REVIEW

The balance-scale task has been studied by developmental psychologists and by modeling researchers using classical and connectionist architectures. In the first section, I review psychological studies using the balance-scale task, including a description of Siegler's (1976) *rule-assessment approach.* Rule assessment figures prominently in both the psychological and modeling literatures. A critical analysis is provided based on problems encountered in the psychological literature. In the second section I review the modeling literature. Both connectionist and symbolic models of the balance-scale task are discussed. In the third section, I argue that although existing models have been successful at capturing the major regularities found in the human data, these data are based on an assessment method that has been the subject of numerous criticisms. Therefore, a new approach to evaluating human and model performance on the balance-scale task is warranted. Despite the numerous problems with rule assessment, it has been the primary method of evaluating the performance of human participants and models, making models based on fitting such data suspect. I suggest an alternative approach -- one that is based on *interpreting* the way a neural network solves balance-scale problems. With this approach it is possible to gain new insights into the problem representation, develop predictions that can then be tested with human participants, and provide alternative methods that can be used to re-evaluate human and model performance.

Psychological Studies Using the Balance-Scale Task

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

Siegler's Model and Rule-Assessment Approach

Siegler (1976) modified the balance apparatus so that there were four equidistant pegs on each side of the fulcrum. For some problems, weights were placed on only one peg on each side of the fulcrum (i.e., uni-peg problems). For others, a number of weights could be placed on two pegs on a single side of the fulcrum (e.g., three weights on the third peg on the left arm and two pegs on the first peg plus three weights on the second peg on the right arm). More recent versions of the task typically do not include multi-peg problems (e.g., Chletsos et al., 1989; Jansen & van der Maas, 1997). Blocks are placed under the arms to prevent the scale from tipping. The participant's task is to predict if the scale will balance, tip to the left, or tip to the right if the blocks are removed. Typically, participants do not receive feedback to ascertain whether a prediction was correct or not (e.g., Aoki, 1991; Chletsos et al., 1989; McFadden, Dufresne, & Kobasigawa, 1987; Siegler, 1976). The correct response to any balance-scale problem can be determined by calculating *torque* (i.e., mass x distance) for each arm and comparing the values.[1]

Siegler (1976, 1978, 1981) developed a rule-assessment methodology to study developmental sequences for various Piagetian tasks using the balance scale as his reference task. The Piagetian or clinical method relied on children's errors and verbal reports. Siegler's method relied instead upon patterns of correct and incorrect responses. This non-verbal approach, in addition to being less informal and subjective, was intended as a means to reveal competence that might have been missed with Piaget's original approach.

Siegler (1976) hypothesized that children go through a series of stages that are governed by the use of different rules. These binary-decision rules are used to assess the importance of the weight and distance information for each problem (see Figures 1 and 2). The proposed succession of rules is as follows. Younger children consider weight alone when deciding whether the scale will balance or not (Rule 1; see Figure

[1] Torque is also defined as *work*, where work = force x distance. Force = mass x acceleration, where the acceleration term is defined by gravity. For purposes of this discussion, torque is simplified to mass x distance, or the number of weights x the number of pegs. For multi-peg problems, the torques for each peg are added together to determine the torque for each arm.

1, Panel A). At the next level, children focus on weight, but will consider distance information in cases where the weights are equal (Rule 2; see Figure 1, Panel B). Next, children realize the importance of both weight and distance, but there is some confusion when one side has the greater weight and the other side has the greater distance (Rule 3; see Figure 2). Rule 3 performance is usually described as guessing or "muddling through" (Siegler, 1976). Lastly, the child or adult can multiply the distance by the weight and compare the products to determine whether the scale will balance or not (Rule 4; see Figure 2). These rules are meant to represent both the underlying knowledge structures and the strategies or processes used in making a prediction (Siegler, 1978).

Siegler (1976) described several different *problem types* that are defined by the combination of weight and distance from the fulcrum (also referred to as the dominant and subordinate dimensions, respectively). *Balance* problems have equal weights at equal distances. Distance is held constant in *weight* problems, so that the side with the most weight goes down. In contrast, the weight is held constant in *distance* problems, so that the side with the farther distance goes down. *Conflict* problems have a different number of weights and distances on each side of the fulcrum. Three types of conflict problem were defined: *conflict-weight* (the side with more weight at a shorter distance goes down), *conflict-distance* (the side with the greater distance but with fewer weights goes down), and *conflict-balance* (despite the conflict, the scale balances). Given all combinations of weights and distances, there exists one additional configuration not explicitly discussed by Siegler. These are problems in which one arm of the balance has both a greater number of weights at a farther distance from the fulcrum. This configuration will be referred to as the *both-greater* problem type.

The assessment of rule use is determined by testing the participant on a set of balance-scale problems. Typically, the test set consists of four of each of the six problem types outlined above.[2] The task is to judge which arm will tip or if the scale will balance. If a participant is using one of the four rules outlined by Siegler, then a characteristic pattern of performance on the different problem types should emerge. Table 1 includes the accuracy predictions for each of the hypothesized rules. In order to classify an individual as using one of these rules, Siegler (1976) outlined a set of criteria. Overall, 20 out of 24 answers had to fit the profile for a particular rule. Moreover, particular types of errors were required for some rules (e.g., a child using Rule 1 should not only get distance problems incorrect, but get them incorrect by predicting that they will balance).

For each of the problem types outlined above, a particular developmental trend was predicted (see Table 1). Performance was not expected to change for balance, weight, or both-greater problems (i.e., all children should show a high level of accuracy). For distance problems, however, a "dramatic improvement with age" was predicted, as accuracy is expected to jump from none correct to all correct. A

[2] There are variations on the size of the test set. For example, Siegler (1976) used 30 problems (4 of each of the 3 simple problems and 6 of each of the 3 conflict problems). Siegler (1981) used 24 problems (4 of each of the 6 problem types). Other researchers have used tests of 24 items (e.g., Aoki, 1991), 25 items (van Maanen et al., 1989), 36 items (e.g., Chlestos et al., 1989), and even 76 items (Ferretti et al., 1985, Experiment 2).

U-shaped trend was predicted for conflict-weight problems. Younger children should get them correct if they focus only on the weight dimension. Once children take note of distance information, performance should drop to chance level as they try to reconcile the two dimensions. When the relationship between weight and distance is understood, performance improves to near perfect levels. Conflict-distance and conflict-balance problems are expected to show the same pattern. Initially, performance is always incorrect; it then "improves" to chance level as participants attempt to incorporate both dimensions. Again, perfect performance follows an understanding of the relationship between weight and distance.

Evaluation of Siegler's Model

Based on cross-sectional studies, Siegler (1976) found a fairly close match between the predictions as outlined in Table 1, and the performance of 120 participants aged 5 to 17. The majority of participants (89%) could be unambiguously classified. There was a clear trend in the ages of the children and the complexity of rule use. Most of the 5- and 6-year-olds were classified at Rule 1. The performance of 9- and 10-year-olds was consistent with either Rule 2 or Rule 3, and Rule 3 classifications were most common for 13- to 17-year-olds. Rule 4 performance was infrequent (7%).

In a subsequent study, the performance of 96% of participants aged 5 to 20 was consistent with one of the four rules (Siegler 1981). Similar age-related trends occurred. It was found that 3-year-olds and half of the 4-year-olds did not use any rule. The other half of the 4-year-olds and almost all of the 5-year-olds used Rule 1.

Older children (8- and 12-year-olds) received Rule 2 and Rule 3 classifications. At the college level, most participants could be classified at Rule 3 with only a minority classified at Rule 4.

Other researchers have evaluated the rule-assessment approach with the balance-scale task using the standardized procedures outlined by Siegler (Aoki, 1991; Chletsos et al., 1989; Ferretti, Butterfield, Cahn, & Kerkman, 1985; Ferretti & Butterfield, 1986; Jansen & van der Maas, 1997; Larivée, Normandeau, Roulin, & Longeot, 1987; McFadden et al., 1987; Normandeau et al., 1989; van Maanen, Been, & Sijtsma, 1989; Wilkening & Anderson, 1982). With respect to the classification of participants there have been mixed results in replicating Siegler's findings. For example, Aoki (1991) could classify only 38% of participants (Grades 4, 6, and college students) as using one of the four rules. Across three studies and different conditions (e.g., individual interviews versus a paper-and-pencil version), Chletsos et al. (1989) found that 70-85% of participants could be unambiguously classified (aged 8-15 and college students). McFadden et al. (1987) tested 5- and 7-year-olds in two studies and found that 80-90% could be classified as Rule 1 or 2. Normandeau et al. (1989) classified 97% of adolescents (Grades 8-11). Ferretti et al. (1985; Experiment 1) tested children in Grades 1-6 and could classify 83% as using Rules 1, 2, or 3.[3] With respect to age-related trends, the general pattern found by Siegler was replicated

[3] Using a procedure in which an "ideal response pattern" was compared with the observed pattern for each subject, Ferretti et al. (1985) computed a ω^2 statistic. All 99 children initially classified with Siegler's criteria could be classified with the ω^2 statistic. An additional 16 children were classified with the ω^2 statistic (Rules 1-3), leaving only 5 children (4%) unclassified.

by other researchers (e.g., Chletsos et al., 1989; Klahr & Siegler, 1978; Ferretti et al., 1985).

Critique of the Rule-Assessment Approach: Psychological Studies

Siegler's (1976) original article describing the rule-assessment method is a seminal paper in the area of cognitive development. As such, rule assessment has been subject to both praise and criticism (e.g., Chletsos et al., 1989; Ferretti et al., 1985; Ferretti & Butterfield, 1986; Flavell, 1985; Jansen & van der Maas, 1997; Kliman, 1987; Larivée et al., 1987; Normandeau et al., 1989; van Maanen et al., 1989; Wilkening & Anderson, 1982). In this section, I will outline the strengths of the rule-assessment method, and then discuss its drawbacks.

The rule-assessment approach was a significant departure from the Piagetian clinical method, and represents an attempt to integrate Piagetian ideas and tasks within an information-processing framework. Unlike the clinical method, there are standardized procedures for administration of the task, and specific procedures for evaluation of performance on the balance scale and other tasks (e.g., Siegler, 1981). One advantage of this standardization is that it has allowed paper-and-pencil versions of the balance-scale task, which makes it far more cost effective than individual interviews (e.g., Chletsos et al., 1989; Ferretti & Butterfield, 1986).

Contributing to the landmark status of this paper is the elegant manner in which the hypothesized rules or strategies could be verified by a very specific pattern of responses, including chance or above chance performance on certain problem types as well as particular types of errors on other problems types. Moreover, the response profile is based on nonverbal responses (McFadden et al., 1987) reducing the influence of a child's verbal competence or performance. The fact that there are three response alternatives (versus two) makes it easier to infer that the response profile was due to systematic factors instead of chance responding (Larivée et al.,1987). Lastly, there is a good fit between the hypothesized patterns of performance and the data from a number of studies (e.g., Chletsos et al., 1989; Ferretti et al., 1985; McFadden et al., 1987; Klahr & Siegler, 1978; Siegler, 1976, 1981; van Maanen et al., 1989).

Despite the advantages just discussed, the rule-assessment approach has also been subject to a number of criticisms. Cognitive developmentalists are interested in both knowledge structures and processing rules (Klahr, 1992; Klahr & McWhinney, 1998). Consequently, a primary goal of this method was to assess the knowledge structures that underlie performance at different stages of development (Chletsos et al., 1989; McFadden et al., 1987; Klahr & Siegler, 1978; Siegler, 1978; Wilkening & Anderson, 1982). Most criticisms, therefore, relate specifically to either the *underestimation* of children's knowledge about balance concepts or to the *incorrect classification* of participants. These criticisms fall into several related categories, including (a) task demands (e.g., multiple-choice format), (b) the possible existence of alternative rules or strategies, (c) properties of the test set that influence rule assessment, and (d) the "arbitrariness" of the criteria used to classify participants.

Task Demands

The forced-choice format of the balance-scale task has been criticized by some as limiting the responses that can be made, and therefore the argument has been made that rule diagnosis may underestimate a child's knowledge level (e.g., McFadden et al., 1987; Normandeau et al., 1989; Wilkening & Anderson, 1982). Some researchers have used a variant of Siegler's original task called a "construction task," (McFadden et al., 1987) or an "adjustment task" (Wilkening & Anderson, 1982) to demonstrate that children may be diagnosed according to a different rule than when compared to assessment with multiple-choice predictions. Children are shown a balance scale with some configuration of weights and pegs on one arm. The child is then given a particular number of weights and instructed to place them at some point on the other arm such that the scale will either balance or tip to a particular side.

For example, McFadden et al. (1987) tested 5- and 7-year-olds on the construction task and Siegler's prediction task. Almost 70% of the 7-year-olds who were originally classified as Rule 1 used the distance dimension preferentially in the construction task. Although the construction task was *designed* to increase salience of the distance dimension, these results can be used to demonstrate that under different conditions and task demands, different behavior (and underlying knowledge differences) may emerge.

Other Rules

Another common criticism of Siegler's model is that it does allow for the possibility that participants may use other rules or strategies. Rather than underestimating knowledge level, the possibility of other rules may result in the *incorrect classification* of participants. The fact that Siegler's procedure usually results in some unclassified participants has been taken as evidence that the model does not include all possible problem-solving strategies (Normandeau et al., 1989). For example, Chletsos et al. (1989) used three alternate forms of a paper-and-pencil version of the task and found 15-30% of subjects could not be classified. Furthermore, the "no rule" classifications were consistent over time. This result can be taken as evidence that some subjects use other rules in a consistent manner, but that these rules may not be diagnosable using Siegler's method.

The lack of clarity regarding the decision rule involved in "muddling through" (i.e., Rule 3) is also regarded as support for the idea of other rules (e.g., Ferretti et al., 1985; Wilkening & Anderson, 1982). It has been suggested that the high percentage of Rule 3 classifications may mask or conceal the use of other rules for the coordination of weight and distance information (Normandeau et al., 1989; Schmidt & Ling, 1996). The existence of other rules is not necessarily a violation of the rule-assessment method. Other rules cannot be detected, however, using the original scoring criteria or test sets (Jansen & van der Maas, 1997). Several researchers have found evidence for rules or strategies other than the four postulated by Siegler (1976, 1981). These include (a) an additive rule (e.g., Wilkening & Anderson, 1982), (b) a "perceptual muddle through" rule (e.g., Klahr & Siegler, 1978), (c) a qualitative proportionality rule (e.g., Normandeau et al., 1989), and (d) a buggy rule (e.g., van Maanen et al., 1989). These are discussed in more detail below.

Additive rule. Wilkening and Anderson (1982) were the first to suggest the existence of other rules that could be used to integrate weight and distance information. In Siegler's model, only Rule 4 involves a true integration of both types of information (i.e., "true" in the sense of consistent with the torque algorithm). Wilkening and Anderson suggested that children may be able to integrate weight and distance information by using either unweighted-addition rules (i.e., compare $LW + LD$ with $RW + RD$) or weighted-addition rules (e.g., compare $2LW + LD$ with $2RW + RD$).

Wilkening and Anderson found that a small percentage (13.8%) of subjects (6-, 9-, 12-year-olds and adults) originally classified as Rule 2 or Rule 3 users could be classified as using an additive rule based on the adjustment task and a test set designed to discriminate between integration and non-integration rules. Ferretti et al. (1985) found that some participants used an addition strategy (3%). Normandeau et al. (1989) classified participants according to Siegler's criteria and a modified system (to test for three other rules and Siegler's rules). Using the original criteria, 60.6% of the participants were classified as using Rule 3. With the modified criteria, 28.3% were classified as muddling through and 26% were classified as using an adding rule. Using latent class analysis, Jansen and van der Maas (1997) also found evidence for an additive rule.

"Perceptual muddle through" (Rule 3A). Klahr and Siegler (1978) first suggested a variant on Rule 3. Here, participants focus on the larger weight or the larger distance and make a perceptual judgment about which side of the balance will tip. A small percentage of the participants (2.4%) were classified as Rule 3A users by Normandeau et al. (1989).

Qualitative proportionality (QP) rule. Participants who use the QP rule take both weight and distance information into consideration. QP users respond as though a heavy weight at a shorter distance should compensate for a lighter weight at a longer distance. For conflict problems, therefore, QP users predict that the scale will balance. Using a modified set of scoring criteria, Normandeau et al. (1989) found that 3.9% of participants used a QP strategy.

Buggy rule. A "buggy strategy" was proposed by van Maanen et al. (1989): "If side X has more weights and the weights on side X have the smaller distance to the tilting point then shift the weights on side X away from the tilting point until the distances on both sides are equal and remove for every shift on side X one weight on side X" (p. 272). In essence, the conflict problem is reduced to a simple balance or weight problem and is easier to solve with Rule 1 (Jansen & van der Maas, 1997). Evidence for this buggy rule was found in a sample of Grade 7 and 8 students by fitting a linear logistic test model (van Maanen et al., 1989). Jansen and van der Maas (1997) re-analyzed van Maanen et al.'s data using latent class analyses and also found evidence for the buggy rule.

Properties of the Test Set

It has been demonstrated that characteristics of the test problems affect participants' behavior and subsequent diagnosis. There are actually two issues here. The first issue has to do with inter- and intra-individual variability in rule use. Siegler's model has been criticized because of the implicit assumption that a participant uses one and only one rule to solve *all* balance problems (Larivée et al., 1987; Normandeau et al., 1989). The second issue has to do with the sources of variability, especially with respect to the properties of the test items used. While it may be the case that different strategies are selected based on the problem types defined by Siegler (1976), a more serious concern involves the particular *instances* within a problem type.

Ferretti et al. (1985) noticed discrepancies between two studies based on differences in the testing sets (e.g., uni-peg versus mixed uni- and multi-peg problems, three different sets of conflict problems). Ferretti et al. speculated that task characteristics may influence strategy choice. To examine this possibility, Ferretti and Butterfield (1986) manipulated the size of the difference in products (or torque) on each side of the fulcrum. To illustrate, consider the two examples of simple distance problems in Figure 3. In Panel A, the torque difference is equal to one. In Panel B, the torque difference is 12 (i.e., the absolute value of [4 x 1 - 4 x 4]). Ferretti and Butterfield used four levels of torque difference: 1, 3, 12, and 24-30 for simple problems and 1, 3, 5, and 18-24 for conflict problems.[4] A test set of 72 items was used: 16 instances of each of weight, distance, conflict-weight, and conflict-distance (with four items from each of the four levels of torque difference), and four of each of balance and conflict-balance problem types. Fewer balance and conflict-balance problems were used because the torque difference is always zero. A rule classification could be determined for each of the four levels of torque difference.

[4] No explanation was provided for the difference in Levels 3 and 4 for simple versus conflict problems. This issue will be discussed in more detail in Chapter 3.

Ferretti and Butterfield found that, in general, as torque difference (TD) increased, so did the percentage of children with a higher rule classification. That is, as TD increased, the number of children classified as "no rule" or Rule 1 decreased and the number of Rule 4 classifications increased. At the highest level of TD, 22% of Grades 1 and 2, 32% of Grades 3 and 4, and 61% of Grades 5 and 6 were classified as Rule 4 users. These numbers are equivalent to or far greater than the percentage of college students classified at Rule 4 in other studies (e.g., Aoki, 1991; Siegler, 1981).

Ferretti and Butterfield also examined the interaction between TD levels and problems types. Accuracy (i.e., number of correct predictions) increased for all four problem types (weight, distance, conflict-weight, conflict-distance) as a function of TD level. The largest differences occurred on the distance and conflict-distance problems. Smaller differences, as a function of TD level, occurred for the weight and conflict-weight problems, but this result was likely due to ceiling effects.

These findings have important implications for the assessment of rule use. Depending on the test items used, children may be classified at either a higher level (if large TD instances are used) or a lower level (if low TD instances are used). Ferretti and Butterfield suggest that "the accuracy of children's rule-assessment classifications may be questioned because these classifications are assumed to be invariant over *theoretically equivalent* problem sets" (1986, p. 1420, emphasis added).

Arbitrariness of the Assessment Criteria

One of the more problematic issues is that the criteria that Siegler (1976) selected for determining a classification were "*arbitrarily* chosen as evidence that a child was using a particular rule" (p. 493, emphasis added). Typically, 20 out of 24 responses must fit the profile for a particular rule. In addition, 3 out of 4 responses for a particular problem type must also match the profile (Siegler, 1981). Siegler claimed that using these criteria, the probability of assigning a rule to a "random responder was less that 5 x 10^{-9}" (p. 494).

These criteria, however, can vary depending on the size and the composition of the test set (see Footnote 2). For example, in order to determine if a particular number of correct (or incorrect) responses is due to systematic or chance factors, the number of problems must be taken into consideration. For example, consider the difference between using four instances of a problem type (e.g., Siegler, 1981) versus six instances (e.g., Siegler, 1976, conflict problems). Given the response alternatives, there is a 1/3 chance of getting an item correct. Using the binomial formula, the probability of answering any number of questions between none and all correct can be determined (Chletsos et al., 1989; see Appendix A for binomial formula and sample calculation). When four test items are used, the probabilities of getting items correct are as follows: none--.197, one--.395, two--.296, three--.098, and four--.012). Using the standard cut-off value of .05, the only case that is significantly different from chance is 4 out of 4 correct. In fact, the value for none correct (which is predicted for some problem types for Rules 1 and 2) does not differ significantly from chance. The probabilities associated with using six test items are as follows: none--.087; one--.263; two--.329; three--.219; four--.082; five--.016; and six--.001. In this case a correct response on 1, 2, or 3 items is more unambiguously interpreted as resulting from chance factors alone (e.g, chance responding is predicted for conflict problems in assessing Rule 3), whereas 0, 5, and 6 items can be interpreted as due to systematic responding.

Other researchers have criticized the rule-assessment methodology because it "lacks a test of goodness of fit" (Wilkening & Anderson, 1982, p. 224). Jansen and van der Maas (1997) addressed the rarely discussed issue of *statistical* information regarding the fit of the predicted rule models to the empirical data. They noted the difficulty in the use of different items, and different numbers of items, by different researchers: "Applying the standards in an unambiguous way to these different data sets is hardly possible and makes it unreliable to compare the results of different experiments on the balance scale task" (p. 328). Jansen and van der Maas suggested that the use of latent class analysis is a potential solution to "the lack of psychometric models of rule governed behavior" (p. 327). Their analyses revealed the feasibility of the rules postulated by Siegler and the other rules discussed above (i.e., additive and buggy rules).

Summary of Psychological Studies

Siegler's rule-assessment methodology represented a departure from the Piagetian clinical method. The use of different problem-solving rules is determined by a characteristic pattern of performance across a number of different problem types. Despite the advantages of this approach, it has been criticized by a number of researchers. Problems surrounding the possible incorrect classification of participants and their knowledge of balance concepts were discussed. There has been a growing interest in modeling cognitive development (e.g., Elman et al., 1996), and as such, the balance-scale task has emerged as a benchmark problem. In the next section, connectionist and symbolic models of the balance-scale task will be described.

Modeling Studies of the Balance-Scale Task

Balance-scale performance has been modeled using both connectionist and symbolic approaches (e.g., McClelland, 1989; Schmidt & Ling, 1996). Although the focus of the present research will be on connectionist modeling, I will begin with a brief description of early models of the balance scale that used symbolic or production system architectures. In the second section, I will describe connectionist models of the balance scale. In each case the goal was to use a computer model to capture reliable psychological phenomena: (a) stage-like performance, (b) U-shaped performance on conflict-weight problems (i.e., weight and distance are in conflict, but the side with the greater weight tips), and (c) the torque difference effect (TDE) reported by Ferretti and Butterfield (1986). Despite the problems with the rule-assessment method, it has been the primary method of testing the performance of computer models. In the third section, I will address an additional criticism of rule-assessment that has come about as a result of using this method for evaluating network models.

Symbolic Models

The earliest attempts to model the balance-scale task used production system simulations (e.g., Klahr & Siegler, 1978; Sage & Langley, 1983). Production systems are composed of rules (i.e., productions) that take the form of *condition-action* pairs (i.e., similar to *if-then* statements). These productions specify the action that should be carried out if some condition is met. Key properties of production systems include a *production memory* (i.e., where the productions are stored) and a *working memory* (i.e., a representation of the current situation). Learning occurs through a process of self-modification, in which productions are created and/or changed based on previous experience (Siegler, 1998).

Klahr and Siegler (1978) used a separate set of production rules for each of the four stages of development (see Figure 4). As there was no transition between these different rule models, the model is silent with respect to issues of stage-like development, transition mechanism, and the U-shaped trend on conflict-weight problems. Klahr and Siegler postulated that the addition and/or modification of particular productions could account for the transition between stages, but this hypothesis was not implemented in these production systems. The set of models appeared before Ferretti and Butterfield (1986) reported the torque difference effect, so it is not possible to evaluate the model with respect to it. In order to display the TDE, however, the system would need to contain an explicit production to take note of the magnitude of the differences. The TDE has been found in children who do not explicitly know the torque rule, so it is difficult to imagine how this issue could be resolved within a production system framework (but see the discussion of Schmidt & Ling, 1996, below).

Sage and Langley (1983) also developed a production system model of the balance-scale task. This revised model was similar to Klahr and Siegler's, but Sage and Langley suggested a possible mechanism to account for the transition between stages. They posited that the rules might be learned through a process of *discrimination*. The model was given two rules that provided initial behavior, and rules for storing information about failures and successes. New rules were acquired, and then weakened or strengthened based on the success of the predictions. Over the course of training, the overall percentage of correct predictions increased (i.e., there was no evaluation of the different problem types). Although the model never mastered conflict problems or developed a representation of torque (i.e., it did not achieve Rule 4 behavior), it learned a set of rules to make correct predictions despite the incomplete representation of the problem, just as children do.

Newell (1990) used the *Soar* architecture to model the balance scale. Similar to the Sage and Langley model, the Soar model did not exhibit performance consistent with Rule 4. The model learned from only a few instances and there was no rigorous comparison of the output with the human data. No attempt was made to test for the torque difference effect.

The most successful symbolic or rule-based model thus far has been Schmidt and Ling's (1996) model of the balance scale using Quinlan's (1993) *C4.5 machine learning algorithm*. The C4.5 algorithm is a general purpose classification system that generates simple decision trees that classify data that vary along a number of dimensions. For each balance problem, seven attributes were presented: (a) whether the problem was a simple balance problem (yes/no format), (b) the side with the greater weight (left, right, neither), (c) the side with the greater distance (left, right, neither), (d) left weight, (e) left distance, (f) right weight, and (g) right distance (all weight and distance values expressed as integers ranging from 1 to 5). The initial simulation displayed an orderly stage progression of all four rules when tested with Siegler's rule-assessment technique. It also displayed a U-shaped trend on conflict-weight problems. The TDE was evident only during Rule 3 performance.

A modified version of the model was run, but attributes (b) and (c) were changed to continuous values between -4 and 4 representing the right side minus the left side for both weight and distance, respectively. Information about weight and distance was no longer presented in an "all-or-none" manner. Rather, the attribute was presented as a graded representation. Under these conditions, the model displayed the TDE at all stages, and the other reliable psychological regularities. Schmidt and Ling concluded that their model was successful, as the selection of problem representation and learning algorithm resulted in a good match to the human data.

Connectionist Models

A brief introduction to connectionist terminology is provided in Appendix B. There are two main types of connectionist networks that have been used to model balance-scale performance. First, McClelland (1989) used the relatively common, multilayer perceptron. This simulation, and replications of it, will be discussed in the first section. Second, Shultz and his colleagues (e.g., Shultz & Schmidt, 1991; Shultz et al., 1994, 1995) used a cascade-correlation architecture. This generative architecture and its results will be discussed in the second section.

The McClelland (1989) Simulation

McClelland (1989) reported the first connectionist model of balance-scale phenomena. The architecture is illustrated in Figure 5. Although Siegler (1976) used a scale with four pegs on each side of the fulcrum with a maximum of six weights on any peg, McClelland used the representation of a scale with five pegs and a maximum of five weights on any peg.[5] Twenty input units were used: the first 10 input units represented the left and right weight values and the remaining 10 represented left and right distance values. Inputs were segregated such that the weight and distance inputs were connected to different hidden units. That is, the first 10 units (representing weight) were connected to the first two hidden units, and the second set of 10 input units (representing distance) were connected to the third and fourth hidden units. The information was integrated only at the level of the two output units (see Figure 5). A higher activation of the left or right output unit was used to indicate the side that would tip, and neutral activation of both units indicated balance.

The segregation of weight and distance information at the hidden unit level is what McClelland (1989) referred to as the *architecture assumption*. McClelland's model also included an *environment assumption*. McClelland assumed that the average child has more experience with balance scenarios in which distance is not important. Instead of training the network on the entire set of 625 possible five-peg, five-weight problems, two corpuses were developed in which there was either 5 or 10 times the equal distance problems (i.e., simple-balance and weight problems). For

[5] Siegler (1976) and other researchers used problems in which weights could be spread across more than one peg (i.e., multi-peg problems). McClelland (1989) and other modeling researchers (e.g., Shultz & Schmidt, 1991) do not distribute weights across more than one peg (i.e., they only use uni-peg problems).

each *training* epoch, 100 patterns were randomly selected. The network was trained using a standard back-propagation learning algorithm. After each epoch, the network was *tested* on a 24-item test set. Note that during the test phase, no new learning (i.e., weight updating) occurred. The network's performance was classified using Siegler's (1976) rule-assessment method.

Without these two assumptions, McClelland's network learned the problems too quickly, often skipping Rules 1 and 2 (Schmidt & Shultz, 1992). Even with these limiting assumptions, there was a lot of shifting between the use of Rule 3 and Rule 4. In fact, Rule 4 was never reliably established. Although this first attempt did manage to demonstrate some stage-like behavior and was a fairly close fit to Siegler's predictions, McClelland did not explicitly test for the torque difference effect. In a replication of the McClelland model, however, Schmidt and Shultz (1991) explicitly tested and found evidence for the TDE (see also McClelland, 1994). Both the original model and the replication failed to exhibit a strong U-shaped trend on conflict-weight problems. Schmidt and Ling (1996) also reported a replication of McClelland's model using a four-peg, four-weight version and a six-peg, six-weight versions of the problem. Both versions resulted in the model producing Rule 3 and 4 behavior only.[6]

Cascade-Correlation

Shultz and his colleagues (Shultz & Schmidt, 1991; Shultz et al., 1994; Shultz et

[6] Interestingly, Schmidt and Ling used these versions because McClelland used a five-peg, five-weight version "despite the fact that Siegler's . . . data was based on a four-peg, four-weight version of the problem" (p. 221). As discussed, Siegler used a *four*-peg, *six*-weight version of the problem. This version was not used in the replication of McClelland's model or for the C4.5 model.

al., 1995) modeled balance-scale phenomenon using the cascade-correlation architecture and learning algorithm (Fahlman & Lebiere, 1990). There are important differences between the cascade-correlation architecture and the multilayer perceptron used by McClelland. Cascade-correlation is described as a *generative algorithm* (Mareschal & Shultz, 1996) or one that "constructs its own network topology as it learns" (Shultz et al., 1994, p. 57). There are no pre-established hidden units in this architecture. The input and output units are fixed, but the hidden units are created and installed by the learning algorithm. To begin with, the architecture is a simple *pattern associator*,[7] with only input and output units (see Figure 6, Panel A). Each input and output unit is connected with a modifiable weight. In this case, there are four input units, representing left weight, left distance, right weight, and right distance. The inputs are integer values between one and five representing the weight and distance values. Shultz and Schmidt also used a maximum of five weights on any one of five pegs. There were two output units. Left and right activation were indicated by a positive value on the tip side and a negative value on the non-tip side. Balance was represented as neutral values on both units.

Initially, the network was trained without hidden units on a sample of 100 balance problems with a 0.9 bias for selecting equal distance problems. The recruitment of hidden units was completed one at a time. When the reduction of error asymptotes, the network trains and evaluates a pool of hidden units. While the hidden

[7] A pattern associator makes an arbitrary mapping between inputs and outputs (Bechtel & Abrahamsen, 1991).

units are being trained, the rest of the network is effectively disabled. When the output of one unit correlates with the residual error signal of the last training epoch, this *candidate* unit is then connected to the active network. The weights for the connections between the input and new hidden unit are frozen and do not change with subsequent training (the connections from the hidden unit to the output units do change with learning). The rest of the network is reactivated and the training continues, again until there is no further reduction of error. If error does not reduce with further training, another hidden unit is installed (see Figure 6, Panel B).

Like the McClelland model, the cascade-correlation model described here needed a bias in favor of training problems with equal distance. Without this training bias, the network immediately went to Rule 3. An additional assumption about the learning environment was used. To simulate the child's gradually changing environment, Shultz and Schmidt used "expansion training," where one new pattern was added to the training set every epoch. McClelland's architecture assumption (i.e., the segregation of weight and distance information), however, was not implemented in this model.

Sixteen "computer subjects" were used in the Shultz and Schmidt experiment (i.e., the simulation was run 16 separate times with different initial weight states). After each training epoch, the network was tested on 24 randomly selected test patterns. Rule diagnosis was similar to that used by Siegler (1976) for human subjects. There were four of each of the six problem types. Each of the four instances of the non-balance problem type represented one of four levels of torque difference (i.e., 1, 2-5, 6-9, 10-20). With respect to the order of rules, 11 computer subjects progressed through all four stages in the appropriate order. Two subjects progressed through the first three rules. One subject used only Rules 1 and 2, one skipped Rule 3, and another skipped to Rule 4 and then regressed to Rule 3. In terms of hidden unit recruitment, nine of the computer subjects recruited one hidden unit, six recruited two hidden units, and one recruited three hidden units. In half of the cases, hidden unit recruitment was associated with a quick progression from one rule to the next.

In a second experiment, training and testing occurred in exactly the same way as described above. However, errors were recorded for test problems at the four levels of torque difference. An ANOVA conducted on the error signals for the middle and last epochs indicated a main effect of torque difference level, such that the larger the torque difference, the smaller the error. Not only did the model go through the stages in the appropriate sequence, it showed use of Rule 4 for most of the computer subjects. There were also instances of stage skipping, developmental regressions, and not achieving Rule 4. These same patterns are found with human subjects (e.g., Chletsos et al., 1989; Siegler, 1981).

Shultz et al. (1995) reported a second cascade-correlation model of the balance scale. Rather than biasing the training environment toward weight information, the internal state of the network was prestructured so that preferential treatment would be given to the weight information.[8] The entire set of 625 possible problems was

[8] This was done by pretraining the network with equal-distance problems. This approach is easier to implement than, for example, supplying connection weights by hand (Shultz et al., 1995).

presented during training. The network displayed the regularities found in human performance, including the use of all four of Siegler's rules in the correct order, some stage skipping and regressions, U-shaped development on conflict-weight problems, and the TDE.

Critique of Rule-Assessment: Modeling Studies

Siegler's (1976) rule-assessment approach has been used to assess the performance of the connectionist models described in the previous section. An additional critique of the rule-assessment method has come to light as a result of these studies. That is, a set of responses can be classified as either Rule 2 or 3, depending on the scoring priority used. Shultz et al. used four scoring variations and found variability in rule assessment. Likewise, two different patterns of stage progressions resulted in Schmidt and Ling's (1996) simulations depending on the order in which the assessment methods were applied. Some regressions occurred when one assessment order was used, but not with the other order. Shultz et al. (1994) noted that this diagnostic ambiguity has not been acknowledged by psychological researchers.

Summary of Modeling Studies

Overall, the connectionist simulations of balance-scale performance have been more successful than the older symbolic models with respect to providing a good fit to the data reported by Siegler (1976). The connectionist models and the C4.5 symbolic model of Schmidt and Ling (1996) have been succcssful at capturing the major regularities of the human data, including stage-like or qualitative shifts in performance.[9] These stages were exhibited by systems that change in a gradual, continuous manner. The U-shaped trend present in the human data, and the torque difference effect were also captured in many of the models discussed. In the next section, I will use the literature reviewed in the first two sections and the various critiques of the rule-assessment method to support my argument that an alternative approach to studying the balance-scale task is needed.

Data Fitting versus Network Interpretation

To date, modeling approaches have been done in the service of fitting the psychological data. For each of the models I have discussed, one of the goals of the researcher was to use an approach analogous to the rule-assessment method to produce network data that conforms to the pattern of results reported by Siegler (1976) and Ferretti and Butterfield (1986). For example, Schmidt and Shultz (1992) replicated the McClelland (1989) model and examined a variety of parameters to determine which variables and interactions produced the most "psychologically realistic" results. They kept the learning rate, momentum, and range of random start connections constant, but manipulated *bias size* (i.e., the number of extra equal-distance problems in the corpus), *subset size* (percentage of the corpus sampled for training), *weight updating* (batch versus continuous), and network *architecture* (segregated versus nonsegregated). The specific results are not at issue, but overall,

[9] Note that a key difference between the C4.5 model and connectionist models is that the C4.5 model does not include any variability in runs--the same solution is generated every time the model is implemented (Schmidt & Ling, 1996). Each run of a connectionist network begins from a random starting state.

any manipulation that slowed convergence resulted in the most psychologically realistic data. What is at issue is that parameters were manipulated in order to fit *one* particular data set, that is, the one reported by Siegler (1976).[10]

This inclination toward fitting data can be amply illustrated. For example, consider Sage and Langley's (1983) evaluation of their model: "The stages through which the system progresses are very similar to those observed in children, so the model provides an explanation of the observed developmental trends" (p. 94). More recently, Schmidt and Ling (1996) commented that, "Regardless of the learning algorithm that one adopts (connectionist or symbolic), the choice of attributes to use is crucial *if the model's output is to match the human data* (p. 211, emphasis added).

What is Wrong With Data Fitting?

Clearly, one goal of modelers has been to fit the psychological data. It would be an overstatement, however, to assert that this was the *only* goal of these researchers. Other goals included determining the conditions under which the observed pattern of performance could or could not be obtained. An example of this is the *environment assumption*. For instance, several modelers needed to bias the training environment in order to observe the progression of rules found in the developmental literature. Another goal was to use computer models to derive novel predictions. For example, Schmidt and Ling (1996) suggested, contrary to several previous models, that "the C4.5 model predicts that the weight and distance dimensions are equally and

[10] I point specifically to Siegler's data, but I also mean to include replications of Siegler as the type of data which modelers have been trying to fit.

symmetrically presented in the natural world" (p. 226). Another prediction from Schmidt and Ling's model concerns the primacy of simple balance problems. Having said that, I want to reiterate that a key goal was to produce the greatest proportion of classifiable networks using the rule-assessment method.

As demonstrated, numerous criticisms have been leveled against rule assessment. These include (a) arbitrary criteria for scoring, (b) assessment varying with the torque difference of the items used, (c) assessment varying with the priority given to the various rules, (d) assessment varying with task demands, (e) scoring criteria that are not diagnostic with respect to other postulated rules, and (f) lack of clarity regarding the "muddle through" stage. These problem, cumulatively, make the interpretation of data from psychological studies either questionable or ambiguous.

Despite these criticisms, a central goal of modeling researchers has been to develop models that fit Siegler's 4-Rule Model. Given the problems encountered in the psychological literature, modeling approaches designed to capture these results will do little to help us understand performance on this task. Elman et al. (1996) nicely summarized this viewpoint:

> " . . . let us be very clear about the important distinction between implementation and theory building. It is of course relatively easy to implement developmental outcomes in connectionist models. Some of the models we have discussed are obviously still at a very early stage of development and go little beyond the implementation of behavioral outcomes." (p. 170)

Essentially, a model is only as good as the data it was designed to fit. In the case of the balance-scale task, modelers have been trying to emulate a pattern of human performance that has been collected with a method that has shown to be problematic. Moreover, given the dearth of alternative assessment methods, this same method has then been applied to evaluate the network models.

Preview of the Present Neural Network Research

Rather than implementing behavioral outcomes, I advocate a novel approach to studying the balance-scale task. The starting point for the research in this dissertation is the assertion that network models thus far have been modeling human data that is open to interpretation because the methods used to collect it have been shown to be problematic. It follows then that the interpretation and evaluation of previous models may also be open to debate.

Currently, new methods have been developed for interpreting network models (e.g., Berkeley, Dawson, Medler, Schopflocher, & Hornsby, 1995; Dawson, 1998; Dawson, Medler, & Berkeley, 1997; Elman, 1990; Hanson & Burr, 1990; McCaughan, Medler, & Dawson, 1999). I will present the results of a connectionist simulation in which the goal was to use these approaches to interpret the network in order to find out how the network solves balance-scale problems. As part of taking an interpretive approach, an analysis of the problem-space was conducted. To preview, this approach has resulted in a new interpretation of the task, new criticisms of the rule-assessment approach, and novel predictions that can be tested with human participants.

CHAPTER 3

EXPLORING NEURAL NETWORKS AND THE PROBLEM SPACE

In Chapter 2, I reviewed the literature on the psychological and modeling approaches to the balance-scale task. Siegler's (1976) decision-tree model makes specific predications about patterns of performance that were borne out in the human data. Likewise, connectionist models provided a good fit to the data by capturing major regularities. The primary method of evaluating both human participants and computer models has been Siegler's rule-assessment approach. The rule-assessment approach was subjected to critical analysis.

Previous connectionist models (e.g., McClelland, 1989; Shultz & Schmidt, 1991) included particular assumptions about, for example, the training environment or the structure of the network. These assumptions, along with the manipulation of other parameters, were done in the service of fitting the psychological data. The psychological data, however, are open to alternative interpretations as they were collected using a method that has been shown to be problematic. Six main criticisms were outlined in Chapter 2. Models that have been designed to emulate these data, therefore, are also subject to alternative interpretations.

Currently, no alternative exists as a method of assessing the performance of either human participants or network models on the balance-scale task. In the first two chapters, I previewed a different way to approach network modeling – one that focuses on *interpreting* artificial neural networks. It is the interpretive approach that is the focus of the present chapter. Here, neural networks are trained on the same task as used with human participants. Instead of manipulating parameters and assumptions to obtain the best possible fit to the rule-assessment data, the fully-trained network is examined in detail to extract the way the network solved the information-processing problem. The main question driving the research becomes: "Can we gain any insight into the task or how individuals solve it by exploring the way a neural network accomplishes the task?"

In the first section, I present the results of a neural network trained to perform balance-scale problems. One network was analyzed in detail to explore questions about how the network solved the task. Through the course of this analysis, many interesting properties of the problem space were discovered. In the second section, I will provide an analysis of the problem space and also of published test sets. As a result of these analyses another experiment with neural networks was conducted. The simulation involved removing a particular type of problem from the training set and then training it first with the same topology as in Experiment 1 and then again with one less hidden unit. The network with fewer hidden units was subjected to interpretive techniques. These results will be discussed in the third section and a chapter summary follows. In combination, the problem space analysis and the interpretation of the networks led to several questions and hypotheses about human performance. These hypotheses and an experimental evaluation follows in Chapter 4.

Experiment 1: Interpreting the Balance Scale Network

In this simulation, neural networks were trained to predict correctly balance-scale problems. Various techniques were then used to determine if the internal structure of the network could be interpreted with respect to questions about how it was solving the balance-scale task. Four main interpretive approaches were undertaken. These will be described in more detail below.

Simulation Methodology

Training Set

The network was presented with all 625 possible five-peg, five-weight problems. Balance-scale problems were represented on 20 input units. The first five units represented left weight and were *thermometer* coded (i.e., a unit was turned on for each of the weights represented; see Figure 7). The second set of five units represented left distance and were *unary* coded (i.e., the unit corresponding to the position of the weights had a value of 1, the others were set to 0). The same pattern was repeated for the remaining 10 units representing right weight and right distance.

Processing Units

The network had 20 input units (see Figure 7). The network was fully connected in that every input unit was connected to every hidden unit (i.e., the input was not segregated as in McClelland, 1989; cf. Figure 5). The network required four hidden units. In pilot tests, this was the minimum number required to obtain reliable convergence. All hidden unit processors were *value units* (Dawson & Schopflocher, 1992). Value units use a Gaussian activation function (minimum of 0, maximum of 1, standard deviation of 1). Value units were selected because the primary goal of this research was network interpretation. Based on previous research, value units have been shown to be particularly interpretable (e.g., Berkeley et al., 1995; Dawson &

Medler, 1996, Dawson, Medler, & Berkeley, 1997; Leighton, 1999; Leighton & Dawson, 1999; McCaughan, Medler, & Dawson, 1999; Medler, 1998; Medler, McCaughan, Dawson, & Willson, 1999).

Two output units were used. Activation of either the left or right output unit represented the corresponding side of the balance scale that would tip. Balance problems were represented by zero values on both units. The output units were also value units.

Training the Network

The network was trained using the generalized delta rule designed for use with the value unit architecture (Dawson & Schopflocher, 1992). The learning rate was set at 0.005, and no momentum term was used. Connection weights were randomly set in the range of ± 0.1 and the biases of all value units (i.e., the mean for the Gaussian for each unit) were initialized to zero. In pilot tests, these parameter settings resulted in reliable convergence. The criterion for a "hit" was set at 0.01 (i.e., the activation had to be greater than or equal 0.90 when the desired output was 1, and less than or equal to 0.10 when the desired output was 0). Pattern order presentation was randomized every epoch. Network connections and biases were updated after each pattern presentation. The network converged after 4120 epochs of training (i.e., the number of epochs required before a hit was recorded for every pattern).[1]

[1] With the parameter values outlined above, the network was run 20 times. The average sweeps to convergence was 4092.1 (*sd* = 120.9). Following the analysis of Network 4120, other networks were evaluated to verify that the results were replicable.

Analysis of the Network

A variety of methods were used to analyze the balance-scale network, including (a) an analysis of the connection weights, (b) a banding analysis of the hidden unit activities, (c) the identification of possible functions, and (d) a cluster analysis of the hidden unit activities. One network was subjected to all analyses.

Approach 1: Analyzing the Connection Weights

Other researchers have examined the connection-weights between processing units as one method of exploring the structure of neural networks (e.g., Hinton & Sejnowski, 1986). For example, *Hinton diagrams* have been used to illustrate the strength and the direction (i.e., positive or negative) of connection-weights.[2] Each connection-weight is plotted as a box. The size of the box represents the magnitude of the connection-weight (zero, small, medium, large) and the color of the box represents the direction (black for negative values, white for positive values).

There were 80 connections in total between the 20 input units and the four hidden units. Instead of using Hinton diagrams, the strength and direction of the connection-weights have been plotted in a standard Cartesian coordinate system (see Figures 8 and 9). The value of the connection-weight is plotted on the ordinate and individual cases are plotted on the abscissa.

First, consider the connection-weights between the input units representing weight information and the hidden units. Figure 8 (solid bars) shows the 20

[2] The connections between units will always be referred to as "connection-weights" or "connections" in order to distinguish them from the weight value (i.e., mass) represented in balance-scale problems.

connection-weights between Input Units 0 though 4 (the five units representing left weight) and the four hidden units. From left to right, the first five solid bars represent the connections to hidden unit 0 (H0), the next five to hidden unit 1 (H1), and so on. Figure 8 (striped bars) shows the 20 connection-weights between Input Units 10 though 14 (the five units representing right weight) and the four hidden units.

Recall that weight was thermometer coded (i.e., a unit is turned "on" for every weight value), therefore, the first input unit for each group of five will always be on (i.e., because there is always at least one weight on each side of the fulcrum). Thus, the connection-weight values for the first case in each of the four groups are similar (i.e., all are near zero: 0.2, -0.003, 0.2, and 0.07 for connections from left-weight units going to the four hidden units (solid bars), and 0.2, 0.1, 0.1, and 0.2 for connections from right-weight units to the four hidden units (striped bars).

One noticeable feature of these connection plots is the symmetry along the horizontal axis. Imagine that for a particular pattern, the left and right weight values are the same (temporarily ignore the distance dimension). As can be seen by comparing the solid and striped bars, the connection-weights to a particular hidden unit cancel each other out, as corresponding values are either positive or negative. If there were two weights on each side of the fulcrum, for example, the value of the connection-weights to H0 for left and right are -.32 and +.32, respectively. Similarly, the left and right connection-weight values to H1, H2, and H3 are +/-1.45, +/-.31, and +/-2.8, respectively. Therefore, when the left and right weights are equal, the signals to the hidden units sum to zero.

Now imagine a difference in weights on each side of the fulcrum (e.g., *LW* = 4, *RW* = 2). In this case the connection-weights for the first two input units cancel each other out as described above, but for the two additional inputs on the left, a positive signal would be transmitted to H2 and H3, and a negative signal to H0 and H1 (note that the opposite pattern would occur if the difference was in favor of the right side). In summary, when there is no difference between the left weight and the right weight, the connection-weights modify the signal from the input units such that all four hidden units receive neutral information. When there is a difference in the values for left weight and right weight, the connection-weights transform the signals so that different hidden units receive incoming signals that are either positive or negative (and that differ in the magnitude of the signal).

Figure 9 shows the connection-weights between the input units representing left and right distance. The 20 connection-weights between Input Units 5 though 9 (the five input units representing left distance) and the four hidden units are represented by the solid bars. The 20 connection-weights between Input Units 15 though 19 (the five units representing right distance) and the four hidden units are represented by the striped bars. Again, there is a symmetry along the horizontal axis. If the distance values on the left and the right are the same, the corresponding connection-weights cancel each other out (one being positive, the other being negative). For example, consider the case where the left and right distance is 1. The connection-weight values for left and right distance are +.53 and -.40 to H0, +2.0 and -2.0 to H1, -.37 and +.44 to H2, and -3.8 and +3.8 to H3. Therefore, when distance is the same on the left and right sides of the fulcrum, the four hidden units would receive signals that sum together to equal zero (or near zero, in the case of the small differences for connection-weights to H0 and H1). Note that we do not have to consider the other units in the group of five that represent either left or right distance. Recall that distance is unary coded, so that the other input units are set to zero. The hidden units will not receive information from those units, as any connection-weight multiplied by zero is zero.

Now consider a difference in distance on each side of the fulcrum (e.g., *LD* = 1, *RD* = 5). As above, the connection-weight values for left distance are +.53 to H0, +2.0 to H1, +.44 to H2, and -3.8 to H3. The connection-weights for right distance are +.34 to H0, +1.4 to H1, -.26 to H2, and -2.6 to H3. Unlike the case when distances are equal, the hidden units receive positive or negative signals from the input units representing distance information. In this case, information from the input units representing distance are passed through connection-weights that either amplify or attenuate the signal. H0 receives two moderately positive signals (.53 plus .35). H1 receives two strongly positive signals (2.0 plus 1.4), H2 receives a signal near zero (+.44 plus -.26), and H3 receives a strongly negative signal (-3.8 plus -2.6). To summarize, when there is a difference between the left and right distance values, the hidden units receive either positive or negative signals, but when the distances are equal, very little information is added to the incoming signal received by any particular hidden unit.

The eight connection-weights between the four hidden units and the two output units are plotted in Figure 10. For any given hidden unit, the connection-weights between it and the two hidden units are large to one output unit and small (relatively speaking) to the other output unit. More specifically, the connection-weights from H0 and H1 are large going into Output Unit 0, and small going into Output Unit 1. The connection-weights from H2 and H3 are large going into Output Unit 1, and small going into Output Unit 0.

Recall that when there was a difference between left and right weight (with more weight on the left side), positive signals were transmitted to H2 and H3, and negative signals were sent to H0 and H1. If we temporarily ignore the processing that the hidden units do we see that the connection-weights would take the positive values at H2 and H3 and send strong signals to the right output unit and weaker signals to the left output unit.[3] The negative signals to H0 and H1 would be scaled to send stronger signals to the left output unit and weaker signals to the right input unit. Clearly, there is a general pattern such that the connection-weights between the hidden units and the output units continue to transmit information about whether there is a difference between the left and the right sides. The story is incomplete, however, because we know that in addition to summing incoming information, processing units (whether hidden or output) also compute a level of activity (i.e., the activation function described in Appendix B).

[3] "Strong" and "weak" are used here in a relative way only for purposes of a general description. Looking closely at Figure 10, some of the connection-weights are actually near zero.

In summary, when the network converges on a solution to every problem in the training set, the connection-weights are configured such that they modify any input pattern so as to discriminate the differences between left and right, for both weight and distance. That is, there is evidence that the network has the ability to integrate the weight and the distance information on each side of the fulcrum. This analysis of the connection-weights provides some insight into how the network has learned to classify patterns as left, right, or balance. To get a more complete picture, however, the activity of the hidden units must also be analyzed.

Approach 2: Banding Analysis

A second way of exploring the structure of a neural network is to analyze the hidden unit activities. Previous researchers (e.g., Berkeley et al., 1995; Dawson & Medler, 1996, Dawson et al., 1999, Leighton, 1999; McCaughan et al., 1999) have found marked *banding* of hidden unit activities when they are graphed on a *jittered density plot*. An example of a jittered density plot and the banding of hidden units is illustrated in Figure 11. The jittered density plot is 1-dimensional scatterplot: hidden unit activity is plotted on the abscissa and the data points are plotted randomly on the ordinate so that they do not overlap.

Banding facilitates an interpretation of the network's algorithm, that is, how the network performed the input-output mapping. Typically, banding is taken as evidence for *feature detection*. That is, each band is associated with input patterns that have some feature in common, and this particular feature assists in the classification of that pattern as a particular output. For example, Berkeley et al. (1995) trained a value unit network on Bechtel and Abrahamsen's (1991) logic problem. Nine out of the 10 hidden units showed banding. Particular types of problem would fall into the same bands. For example, ***modus ponens*** problems clustered together, as did ***modus*** *tollens* problems. By analyzing the network in this way it was possible to determine which features the network used to solve the problems. Next, I will describe the procedure for conducting a banding analysis.

Recording hidden unit activities. The activities of the hidden units were recorded after the network had converged on a solution to every problem in the pattern set. The recording of hidden units is called *wiretapping*, as the procedure is analogous to single-cell recording in the brain (Dawson, 1998). Once the network has converged (i.e., has correctly learned to solve the problem it has been presented it with), *each* pattern in the training set is presented to the network one more time. For every pattern, the hidden units compute a particular level of activity. The internal level of activity for each hidden unit is then recorded for every pattern in the training set. In the case of the balance-scale task, for each of the 625 patterns, the activations of the four hidden units were recorded (resulting in a 625 x 4 data matrix).

As can be seen in Figure 12, banding was *not* evident. There were regularities, however, in the "smearing" of hidden unit activities. For example, for hidden unit 0 (H0), all the right patterns had low activation, left patterns had high activation, and balance patterns were in between. The reverse pattern was found on hidden unit 2 (H2). For H1 and H3, there was a mixture of left, right, and balance problems across the range of hidden unit activity values. Some of the "loose" bands on H1 and H3 contain the same types of problem (e.g., all left patterns), but these alternate across the range of values.

Although the analysis of the hidden units did not reveal "feature detectors," the regularities in the hidden unit activities may reflect an integration of information. In fact, the lack of feature detectors supports the same general notion that was found in the analysis of the connection-weights: that the network is performing some type of integration of weight and distance information. Neural networks are also able to classify patterns without detecting discrete features (Dawson, 1998). One way that neural networks classify patterns is to detect the features that allow a classification. A second way involves the calculation (or approximation) of some function. For example, Medler and Dawson (1994) trained a neural network to simulate reaching for an object. The problem involved a crab-like robot that has two "eyes" that rotated and an arm that reached for an object placed in front of it. The input to the network was the angle of rotation of the eyes required to fixate on the object. The model was trained to output two values: the angle of the shoulder joint and the angle of the elbow joint. These values were continuous and represent the bends in the arm required to make contact with the object. The network must approximate two functions: the angle of both the elbow and the shoulder joint is a function of the angle of rotation of the left and right eyes. Next, the idea that the network was calculating a function was explored.

Approach 3: Identification of Possible Functions

The regularities in the jittered density plots (see Figures 12) and the regularities in the plots of the connection-weights (see Figures 8, 9, and 10) together are suggestive of the hypothesis that the network was classifying problems as left, right, or balance via *function approximation*. The approximation of a function seems plausible given that a mathematical equation can be used to determine the outcome for balance-scale problems. Given the hypothesis of function approximation, the next step was to identify the function.

Given that the network solved all problems correctly it seemed necessary to assume that all four dimensions were integrated in *some* manner. A few functions were likely candidates, including the torque algorithm and the unweighted additive ($[RW + RD] - [LW + LD]$) and weighted additive (e.g., $[2RW + RD] - [2LW + LD]$) integration rules suggested by Wilkening and Anderson (1982). The hidden unit activities were plotted as a function of the torque equation (see Figure 13) and as a function of an unweighted additive equation (see Figure 14).[4]

These two figures are very similar. The discontinuities in the lines in Figure 13 are a result of the gaps in the values of the torque equation on the abscissa. In the case of both the multiplicative and additive equations, the network responded in a particular way to problems depending upon the size of the difference between the left and right sides. When there were large differences in the amount of "information" (i.e., torque difference or additive difference), a single hidden unit responded (i.e., H0 and H2 for left and right problems, respectively). When the difference was small, a set

[4] The weighted equations were also plotted, but the general pattern was the same as illustrated in Figures 13 and 14. Only the unweighted additive equation will be considered further.

of hidden units was required to make the correct prediction. The cross-over point for H0 (the left detector) and H2 (the right detector) occurs at zero in Figures 13 and 14. The highest activations for H1 and H3 are at the zero point on both figures (i.e., balance).

This situation can be thought of as analogous to the overlapping receptive fields in the visual system. This analogy is illustrated in Figure 15. In panel A, with non-overlapping receptive fields, an object is detected by a large receptive field, but information about the precise location cannot be detected. With overlapping receptive fields the activation of *three* sensors provides the additional information required to locate an object more precisely in space. Similarly, fine discriminations in the network were required when the difference between the two sides was very small. The "overlapping" activity of the hidden units signals that the difference is small--but provides enough information to make the appropriate response. This is called *coarse coding* (Hinton, McClelland, & Rumelhart, 1986).

In order to determine if the network might, in fact, be approximating a multiplicative or an additive function, the *net input* for each hidden unit was computed for each of the 625 balance problems. The procedure was as follows. For each hidden unit, there are 20 weighted connections between it and the 20 input units. From the topology of the converged network it is possible to determine the value of these connection weights (see also Figures 8 and 9). Using matrix algebra, one can determine the net input to a hidden unit for every possible input pattern by summing together the products obtained by multiplying the input patterns times the connection weights. More precisely, this is the net input function that involves the summation of incoming information (see Appendix B).

For each of the four hidden units the net input for each pattern was plotted as a function of the torque equation (Figure 16) and as a function of the additive equation (i.e., [*RW* + *RD*] - [*LW* + *LD*]; see Figure 17). The associated correlation coefficients are shown in each panel. For three of the hidden units there was a very high correlation between the torque equation and the net input, and an almost perfect correlation between the additive equation and the net input. For hidden units 2 and 3 there was one outlier. This pattern is the simple balance problem with five weights on the fifth peg on each side of the fulcrum. When this outlier is removed, the correlation between net input for hidden unit 2 changes to -.92 in the case of the multiplicative equation, and -.97 in the case of the additive equation (i.e., the absolute value is the same for all four hidden units for both equations).

The evidence for an additive or a multiplicative function is not clearly distinct (cf. Figures 13 and 16 with Figures 14 and 17). It is unlikely, however, that the neural network was actually performing multiplication. Multiplication is not a *primitive operation* of the processing units. Recall that an individual processing unit (i.e., either a hidden or an output unit) has a net input function, and that the most common type (and the type used in this network) is one that simply *sums* the information from all incoming units. This point is illustrated in Figure 18. Although an outgoing signal is multiplied by a connection weight, the individual processing unit sums together the information about weight and distance. The processing unit cannot, therefore, directly multiply the weight and distance information. It is more likely that the network has learned the mapping between input patterns and output responses by *approximating* the multiplicative rule.[5]

Thus, the evidence so far points to the idea that the network has learned how to integrate the weight and distance information, and that it might be solving problems via function approximation. Next, I present the results of the cluster analysis approach to network interpretation.

Approach 4: Cluster Analysis of Hidden Unit Activities

The hidden unit activities were subjected to a *k-means cluster analysis*. The k-means algorithm is an iterative procedure that involves three main steps: (a) the data set is partitioned into *k* different clusters; (b) the centroids of each cluster are calculated; and (c) each data point is assigned to the cluster that has the nearest centroid. The process is repeated until no data point is moved to a different cluster. The heuristic for selecting *k*, the number of user-defined clusters was as follows. Given a converged network (i.e., one that has correctly learned the task), we know that there has been a correct mapping of training patterns, hidden unit activities, and output responses. The heuristic stopping rule takes advantage of this mapping. That is, we take the smallest number of clusters such that each contains unique output responses. In the case of the balance-scale task, all members of a cluster would have to be patterns with the same output prediction -- either all left, all right, or all balance (see Dawson, Willson, McCaughan, & Medler, 1999, for more details).

[5] This issue will be discussed in more detail in the analysis of the problem space, below.

Using this stopping heuristic, the smallest number of clusters that met the criterion was seven. The division of the 625 patterns into clusters can be found in the first two columns of Table 2. Three clusters contained *left* problems, three contained *right* problems, and one cluster contained all *balance* problems.

Given there are seven types of balance problems (see Table 1) and the seven cluster solution, an obvious question was whether there was a mapping between cluster membership and the problem types described by Siegler (1976). Recall that these problem types are essential for differentiating the use of the different rules in Siegler's model. Other researchers have used cluster analysis as a tool for analyzing the structure of a converged network (e.g., Dawson et al., 1999; Elman, 1990; Hanson & Burr, 1990; McCleod et al., 1998). Often, there is a mapping between the characteristics of the problem and the clusters. For example, the *mushroom problem* is superficially similar to the balance-scale task. In this benchmark machine-learning task a decision must be made about whether a mushroom is edible based on its features (e.g., color, odor). Using binary decision trees over 8000 different mushrooms can be classified as either edible or poisonous. A neural network can be trained to do this task and based on the cluster analysis, groups of features that predict edibleness cluster together (Dawson et al., 1999). So for example, all patterns with an anise odor fall into the same cluster (i.e., all mushrooms with an anise odor are poisonous).

In the present case, however, there was no clear mapping of the seven problem types to the seven clusters. This point is illustrated by comparing Figures 19, 20 and

21. The *pattern space* for the training set cannot be shown properly in four dimensions, so these figures illustrate the pattern space in two dimensions. The dimensions were collapsed by plotting left torque (i.e., *LW* x *LD*) by right torque (i.e., *RW* x *RD*). The entire problem space (or *torque difference space*) is shown in Figure 19.[6] The 625 problems are shown with a slight random jitter to show problems that overlap in the space. The carving of the torque-difference space by standard problem types (Figure 20) reveals complete overlap for weight and distance problems (Panels B and C) and complete overlap for conflict-weight and conflict-distance problems (Panels D and E). That is, they cover the exact same points in the problem space. Moreover, these four problem types overlap with each other and with the both-greater problems. In Figure 21, the pattern space is carved by the cluster analysis into non-overlapping sections (i.e., there is no overlap in the torque difference space).

Although Siegler (1976) has suggested that the balance-scale task (like the mushroom task) could be solved via binary decision trees (see Figures 1 and 2), balance problems do not have many discrete (or nominal) characteristics. In fact, when we consider the features of balance problems, there is only one discrete characteristic: *problem type*. The other features that differentiate the problems have to do with *continuous* characteristics,[7] such as the weight and distance dimensions, or the difference in weights, distances, sums, or torque for each arm. The results of the

[6] A torque difference space seemed the appropriate way to consider the task in two dimensions, as the mathematically-correct predictions map onto regions of the space. In Figure 19, all balance predictions fall on the diagonal, and all left and right predictions fall to the left and right of the diagonal.

[7] “Continuous” is used loosely here as numeric and ordinal – to contrast with discrete.

cluster analysis can be used to support the idea that the network did not solve balance-scale problems via pattern classification based on the *type* of problem, but rather, based on *where* the problems were in the problem space.

Consider the mean hidden unit activities for each of the clusters presented in Table 2 (also see Figure 22). Given that the cluster analysis is based on differences in these values, it is not surprising that hidden unit values differ across cluster. What is noteworthy, however, is that there is a *pattern* of hidden unit activity that characterizes each cluster. For example, hidden unit 0 (H0) is associated with *left* patterns. Cluster 1 includes the patterns with large torque differences (TD) and strong activation of H0. Cluster 7 includes patterns with a medium level TD and a medium-level activation of H0. The same pattern holds for *right* patterns, with H2 having strong and moderate levels of activation for patterns in clusters 2 and 5, respectively.

A slightly different pattern emerged for patterns with low TD. For left and right patterns with low TD, there were rather high activations for H1 and H3 respectively, along with low levels of activation distributed across the other two hidden units. Similarly, balance problems (i.e., TD = 0) were associated with distributed hidden unit activities. In general then, the network was sensitive to differences in torque. Recall that the torque difference effect was found by Ferretti and Butterfield (1986) in human subjects, and found independently by Shultz and Schmidt (1991) using cascade-correlation networks.

Let us return to the plots of the connection-weights (see Figures 8 and 9). Notice the large differences in connection-weight values for input units connected to H1 and

H3 (cases 1-5 and 11-15). Contrast this with the connection-weight values for H0 and H2 (cases 6-10 and 16-20). This same pattern is evident for left weight (Figure 8, solid bars), right weight (Figure 8, striped bars), left distance (Figure 9, solid bars), and right distance (Figure 9, striped bars). There are extreme differences in connection-weights connecting to the hidden units that are associated with balance and small torque-difference problems (i.e., H1 and H3). These extreme values are not seen in the connection-weights to the hidden units that are associated with large torque-difference problems. The large differences in connection weights ensure that when there is a *small* difference between the left and right values (for either weight or distance), this information will be propagated through the remaining layers of the network.

Discussion of the Network Analysis

The results of the approaches used in analyzing the neural network converge on two important ideas. First, problems were solved by the network based on continuous properties of the problem, not discrete or nominal characteristics. That is, the classification was done via function approximation and this function was most likely an additive equation that included all four dimensions of the task (i.e., *LW, LD, RW, RD*). Problems clustered together based on numerical aspects of the problem (see Figure 22), and not the traditionally defined problem types.

The second idea concerns the importance of the problem space. The integration was tied to the location of problems in the torque difference space. When the difference between the left and the right side was large, a single hidden unit was sufficient to signal the correct response. When the difference was small or zero, the overlapping activity of several hidden units was necessary. Small and large differences in the left and right torque can be visualized by reference to the problem space (see Figure 19). Next, the problem space for balance-scale problems was examined.

Analyzing the Problem Space

Recall that when hidden unit activity was plotted as a function of the additive and the multiplicative equations, the general pattern was the same (see Figures 13 and 14). Likewise, the correlations between net input and both equations was very high (approximately ±.97 and ±.92 for additive and multiplicative, respectively; see Figures 16 and 17). Although I asserted that it was unlikely that the neural network had discovered how to multiply, the similarities between the additive and the multiplicative equations begged for further investigation.

For each of the 625 patterns, solutions were calculated using both the torque rule and the additive rule. Figure 23 shows the scatterplot of the two solutions. The correlation between the torque rule and the additive rule was 0.95. Using an additive heuristic, 573 out of the 625 problems (91.7%) can be solved correctly (i.e., the same solution as would be generated by the torque algorithm). This finding provides insight into how it was possible for the network to approximate the torque algorithm when multiplication is not a primitive operation. This finding also has implications for the psychological literature. There has been debate about the nature of the integration that occurs when an individual is aware of the importance of both dimensions (and yet does not know the torque rule). This simple fact about the nature of the problem itself suggests the need for a reinterpretation of the human literature. Following a discussion of the 8% of problems in which the solutions using the torque and additive rules do not match, the composition of test sets used in the balance-scale literature will be analyzed.

The Match versus No-Match Distinction

What are the characteristics of the balance-scale problems for which there is *no match* between the solutions generated by the torque and additive rules? The 52 problems that do not result in a correct solution when the additive rule is applied are all conflict problems (4 CB, 24 CW, and 24 CD), and all have torque differences between ±4 (see Table 3; a list of all no-match problems appears in Appendix C).

Some researchers have used the no-match problems as a means of differentiating between individuals who used an additive or a multiplicative strategy (e.g., Ferretti et al., 1985). The frequency of occurrence of these items, however, has not generally been acknowledged. There are *six* different types of no-match problems (see Table 3). Of the 24 conflict-balance problems, only four are no-match (torque predicts balance but additive predicts tip). A further distinction can be made. For two of the CB problems, the additive rule predicts that the scale will tip to the side with the greater weight (conflict-balance sum-weight, or CBSW), and for the other two it predicts that the scale will tip to the side with the greater distance (conflict-balance sum-distance, or CBSD).[8] The CBSW and CBSD problem types each constitute 0.3% of the entire problem space.

There are 48 conflict-weight and conflict-distance problems that can be classified as no-match. For 40 of the problems the torque rule predicts the scale will tip but the additive rule predicts balance. Twenty are conflict-weight sum-balance (CWSB), and 20 are conflict-distance sum-balance (CDSB). These 40 problems represent 6% of the problem space. For the remaining eight problems, torque predicts the scale will tip to one side, but the additive rule predicts the scale will tip to the *opposite* side. In the case of conflict problems this implies the side with the greater quantity on the opposite arm. Four of the problems are conflict-weight sum-distance (CWSD), and four are conflict-distance sum-weight (CDSW). Each type represents 0.6% of the problem space.

The implications of the additional classification of problem types can be demonstrated by referring to Table 4. The predictions for three common strategies are shown for each type of conflict problem, separated into match and no-match problems. Shown are the mathematically-correct torque rule, the additive rule, and a "weight only" rule. Recall that participants using Siegler's Rule 1 and Rule 2 consider only the weight dimension when confronted with conflict problems.[9] Sielger's Rule 3 is not listed. The predictions for Rule 3, according to Siegler's description, would

[8] These labels were suggested by Ferretti et al. (1985).

[9] There is a .67 correlation between the solution generated by the mathematically correct rule and the solution generated using weight information alone. That is, 413 out of 625 (66%) problems can be solved correctly using a weight-only rule.

involve random responses for any type of conflict problem.

Notice that the first four types of no-match problems make predictions that are the same for two out the three strategies listed. For example, for CBSW problems, both the additive and the weight-only rules predict that a participant will choose the side with the greater weight. Therefore, it would not be possible to use this type of problem to discriminate between users of these two strategies. The same holds for the CDSW, CWSD, and CWSB problems. Only the CBSD and the CDSB problems can be used to discriminate among the three different strategies. Similarly, the conflict problems in which there is a match between the solutions generated by the torque and the additive rules cannot be used to discriminate between users of different strategies. Again, two of the three strategies predict that a participant using that rule would make the same response.[10]

Given the extended classification of balance-scale problems, I examined the test sets used by previous researchers in order to determine whether the no-match problems were included and if so, which types of no-match problems were included.

Characteristics of Published Test Sets

Given the preceding analysis of the problem space a closer examination of the types of instances that have been used in the test sets used in the psychological and modeling research is warranted. Very few authors have published the items used in their test sets, but I will show that of those available there is a confounding of the

[10] The analysis and predictions outlined in Table 4 do not take the torque difference effect into account. This issue will be addressed further in Chapter 4.

match and no-match types and of torque difference. I analyzed the test sets used by (a) McClelland (1989; reported in Schmidt & Shultz, 1991), (b) Jansen and van der Maas (1997); and (c) Schmidt & Shultz (1991).

McClelland (1989) and Schmidt and Shultz (1991). The test set used by McClelland (1989) was adapted from Siegler's (1981) set of test items. Siegler's exact test set could not be used, as McClelland used a maximum of five pegs and five weights and Siegler used a maximum of four pegs and six weights. This test set was used to evaluate a replication of McClelland's model and appears in Schmidt and Shultz (1991). This test set was also used to assess the performance of the models reported in Schmidt and Ling (1996) and Shultz et al. (1995).[11] Table 5 shows the absolute torque difference (TD) of the non-balance items in the test set.[12]

In this test set there is an obvious confounding of problem type and TD. Consider, for example, the weight problems. As discussed previously, Ferretti and Butterfield (1986) discovered that the larger the torque difference, the more likely it is that an individual will be diagnosed as using a more sophisticated rule. Three of the weight problems had a TD of 8, and one had a TD of 4. Likewise, there was a range of TD values for distance problems (including one problem with a TD of 10). Given these large values and the existence of the torque difference effect, the simple problem types used in this test set may inadvertently bias the testing situation such

[11] Shultz et al. (1995) reported the results of a cascade-correlation model (see Chapter 2). In Shultz's earlier cascade-correlation model (Shultz & Schmidt, 1991), 24 *randomly selected* items were used every epoch so it is not possible to assess the characteristics of their test sets.

[12] Excluding *both-greater* problems. None of the research to date has included this problem type in test sets. It has, however, been used in the training sets in modeling research.

that a participant (or a neural network) will get these problems correct. If an individual or a network is to be classifiable, it is important that the simple weight and distance problems are solved earliest.

Consider the conflict problems in Table 5. There were two no-match problems (one CDSB and one CWSB, both TD = 2). Both were problems in which the torque rule predicts tip to one side but the additive rule predicts the scale will balance. One of the four conflict-balance problems was also a *no-match* problem. In this case torque predicts balance but the additive rule (and the weight-only rule) predicts that it will tip to the side with the greater weight (i.e., CBSW).

In summary, the simple weight and distance problems had large torque differences, and the conflict problems had small torque differences. Moreover, 25% of the conflict problems were no-match problems. McClelland (1989) and Shultz and Schmidt (1991) were not specifically looking for evidence of other strategy use (e.g., additive versus multiplicative rules). Rather, they were looking for evidence of the stage-like progression found in the human data (e.g., Siegler, 1976). The composition of the test set may have inadvertently biased the results such that analogous stages were found in the model. Given that the test set was based on one used by Siegler (McClelland, 1989), this same bias may have affected the results of the original psychological studies. In this case though, the test set may have biased the human results such that they matched the predictions of Siegler's decision tree model.

Jansen and van der Maas (1997). Jansen and van der Maas used a test set of 25 items to evaluate human participants and a replication of McClelland's (1989) model.[13] The number of non-balance items used appear in Table 6. Five of each problem type were used, excluding simple-balance problems. Similar to the previous test set the values of TD were larger for the simple problems than they were for conflict problems. Again, the simple problems were made easier by the fact that instances with much larger values of TD were selected.

For conflict-distance (CD) problems, four were no-match problems (three were TD = 1, one was TD = 2) and one was a match problem (TD = 2). The four no-match problems were all CDSB. As discussed previously, CDSB is one of the types of no-match problem that can discriminate users of three different strategies (see Table 4).

In the set of five conflict-weight (CW) problems, two were match problems (TD's of 1 and 2), and three were no-match problems (two with TD = 1, one with TD = 2). All no-match problems were CWSB (i.e., torque predicts tip to the side with greater weight, but additive predicts balance). The five conflict-balance problems that were used were all match problems (i.e., the same solution would be arrived at using either the torque rule or the additive rule).

Schmidt and Shultz (1991). Following Ferretti and Butterfield (1986), Schmidt and Shultz (1991) developed four separate sets of test items in order to do rule diagnosis such that problem type and torque difference were not confounded. Ferretti and Butterfield defined four levels of torque difference and used items from a set with a maximum of six weights and six pegs (see Appendix D for the number of items at

[13] The test set was originally used by van Maanen et al. (1989), but is published in Jansen and van der Maas (1997).

each level of torque difference for non-balance problems). Schmidt and Shultz adapted this set for use with the five-peg five-weight scenario (see Appendix E).

As mentioned, levels of torque difference (TD) were not confounded as they were in the McClelland and Jansen and van der Maas test sets. For some sets, however, there is a confound with the match/no-match distinction (see Table 7). At the first level (all problems have TD = 1), one of the four CD and two of the CW problems were no-match problems. In this case, they were all problems in which torque predicts the scale will tip in one direction but the additive rule predicts the scale will tip in the *opposite* direction (i.e., one CDSW and two CWSD).

The second level of TD included problems with an absolute TD of 3. In this test set, all four CW problems were match problems (i.e., they can be solved correctly using a torque rule, an additive rule, or a weight-only rule). In the case of the CD problems, two were match and two were no-match problems. The two no-match problems were CDSB (see Table 7).

The same balance and conflict-balance problems were used in each set (i.e., for all four levels of TD). In the case of the conflict-balance problems, three were match and one was no-match. The no-match problem is one in which torque predicts balance but the additive rule and the weight-only rule predicts it will tip to the side with the greater weight (i.e., CBSW). For this type of item, it is not possible to determine if an individual is using an integrative rule or not.

Torque-difference Level 3 (TD = 12 for simple problems, TD = 5 for conflict problems) and Level 4 (TD = 15-20 for simple problems, TD = 10-15 for conflict problems) were not analyzed with respect to the match/no-match distinction because all problems with torque differences of greater than 4 are match problems. As noted in Chapter 2, it is not entirely clear why there was such a large difference in TD for simple and conflict problems that were defined as the *same level* of TD. Neither Ferretti and Butterfield (1986) nor Schmidt and Shultz (1991) provided a rationale for this decision. As can be seen in Appendices D and E, however, it was possible to select comparable values of TD for both simple and conflict problems (and for both the 6 x 6 and the 5 x 5 problem sets). To make matters worse, by using the same or comparable levels of TD as Ferretti and Butterfield, Schmidt and Shultz selected problems in which the TD at the third level for simple problems was *greater than* the TD values at the fourth level for conflict problems (see Appendix E).

Although Ferretti et al. (1985) used no-match problems and noted the distinction between the different types, it is not clear if they did or did not use these types of problems when they controlled for torque difference (Ferretti & Butterfield 1986). The test set was not published in either paper. In Figure 1 of Ferretti and Butterfield, however, one of the four illustrations of conflict-weight problems includes a CWSB problem. Therefore, it is not possible to determine if the no-match problems were used, and if so, if they were counterbalanced within levels of torque difference (for the first and second levels), or for conflict-balance problems used at all four levels of torque difference.

Experiment 2: The "Match Only" Network

Given the analysis of the problem space, and the discovery that a small percentage of problems (8.3%) could not be solved with a general adding strategy, a network was trained on a set of problems that did not include the no-match problems. That is, the network was presented with the 573 problems in which the same solution results from using either the additive or the multiplicative rule. It was predicted that if this small part of the problem space was removed, a neural network would either (a) solve the problem more quickly (as indexed by sweeps to convergence), or (b) solve the problem with fewer hidden units.

Simulation Methodology

The same network topology (see Figure 7) and parameter values were used as in Experiment 1 (i.e., 20 input units, 4 hidden units, 2 output units). All 573 *match* problems were presented to the network. A second network with only 3 hidden units was also run. All other details were the same.

Results

Over 20 runs, the network with four hidden units converged in 103.9 sweeps (*sd* = 39.6; range 79 - 208). Recall that for the network presented with all 625 problems, the mean number of sweeps was 4092.1. When the topology was simplified by removing one of the hidden units, the network converged reliably. Over 20 runs, the average sweeps to convergence was 1938.2 (*sd* = 1361.4; range = 92 - 3033). One of the networks with three hidden units was analyzed in more detail. This network converged in 2929 sweeps.

Banding Analysis

Figures 24 and 25 show the jittered density plots of hidden unit activity for the three hidden units in Network 2929. As can be seen in Figure 24, the same pattern was found for H0 and H2. Although there were distinct bands, all bands with *left* patterns had low activation. All *balance* problems fell into one distinct band following the left patterns. Bands containing *right* patterns followed the balance problems and had the largest activations.

Figure 25 shows the jittered density plot for H1. Although on gross inspection three bands are evident, a closer inspection revealed that the first two bands each contained two discrete "microbands." The overall pattern deviated from that seen for H0 and H2. The two microbands near zero contained a mixture of left and right problems. The next microband contained all balance problems, followed by a microband with all left patterns. The band with the largest activation contained only left patterns.

As mentioned in the analysis of Network 4120 (Experiment 1), there are few discrete features that differentiate balance-scale problems. So what were the "features" of the patterns that fell into the bands? The traditionally defined problem types were not associated with bands (with the exception of bands that contained balance problems). The feature that was shared by band members was the solution to the additive equation. This relationship is illustrated in Table 8 for H1. Values of the additive equation (rows) are crossed with the bands for H1 (columns). In some cases, all patterns with a unique solution to the additive equation fall into a single band (i.e., -2, -1, 0). Larger absolute values of the additive equation fall into band A1.

This general pattern held for H0 and H1 (tables are not shown). These hidden units had more bands, and so there were more cases of unique solutions falling into distinct bands. For example, H0 had 15 bands and there are 17 unique solutions to the additive equation (range -8 to +8) for the five-peg, five-weight version of the task. There was only one band that contained patterns with different solutions (i.e., micoroband A1 contained patterns in which the additive equation resulted in solutions of 6, 7, or 8).

Cluster Analysis

Given the relationship between band membership for single hidden units and the additive equation, a k-means cluster analysis was performed on all three hidden units. The general procedure was outlined in Experiment 1. The minimum number of clusters that could be extracted was six. Again, the four dimensions of the problem must be collapsed in order to view them in two dimensions. The problem space in this case can be thought of as an "additive difference space." The clusters are plotted in this additive difference space in Figure 26. As was seen with the torque difference space (see Figure 21), balance problems fall on the diagonal (cluster 6). Left and right patterns fall to each side.

Discussion of the Network Analysis

As in Experiment 1, there was a carving of the problem space based not on discrete features but on values of the additive function. The network in the present experiment required only three hidden units when 8% of the problems were removed from the training set. From the analysis of the problem space, it was known *a priori* that the 573 problems could be solved using an additive equation. The match-only network provides empirical support for the ease with which this large proportion of the problem space can be solved.

The purpose of Experiment 2 was to use neural networks to further explore a particular characteristic of the problem space – the distinction between problems that can be solved using either an additive heuristic or the torque algorithm (i.e., match problems) and those that can only be solved correctly using the torque rule. Experiment 2 demonstrated that the minimum architecture required for all 625 problems could solve the problem in approximately one-fortieths the sweeps when the no-match problems were removed from the training set. One hidden unit could be removed and the network converged reliably. These results converge on the idea that there is something particularly "difficult" about the no-match problems and that there are potential implications for human performance.

Chapter Summary

In the first experiment, a neural network was trained to perform correctly all balance-scale problems. One network was subjected to in-depth analysis. A cluster analysis on the hidden unit activities demonstrated that the network was not using discrete features of the problems to make predictions. Rather, problems clustered together based on where they fell in the problem space. This finding and the fact that balance-scale problems can be solved via a mathematical equation led to a consideration of a function approximation solution. An additive rule seemed a likely candidate for the function being approximated. This assertion is based on the idea that multiplication is not a primitive operation of the neural network, and further supported by the problem space analysis that revealed that 91% of problems could be solved correctly using an additive heuristic.

The analysis of the problem space motivated an examination of the test sets that have been published. This analysis revealed the confounding of problem type with torque difference and the match/no-match problems. Moreover, previous test sets were confounded in such a way as to potentially bias the results in favor of fitting the predictions of Siegler's model. A second simulation was run without the no-match problems, demonstrating that this small part of the problem space made the problem more difficult for the neural networks and potentially for human participants as well. The analysis of the published test sets revealed that these no-match problems typically are present.

In previous work on the balance-scale task (e.g., McClelland, 1989; Shultz & Schmidt, 1991; Siegler, 1976) the problem types defined by Siegler are essential for determining the rule used to solve balance problems and in the case of human participants, for characterizing knowledge level. The results of the network interpretation and problem space analysis can be used to demonstrate that a complete picture of performance on this task will not come from examining a small sample of the six, traditionally-defined problem types. Next, the implications of these analyses will be drawn out with respect to a new approach to evaluating performance on the balance-scale task with human participants.

Preview of the Psychological Research

In Chapter 4, I will present the results of a study in which human participants were presented with the entire set of balance-scale problems (as was done with the neural network in the present chapter). This is a departure from the usual procedure with the balance-scale task. As reviewed in Chapter 2, participants typically have been presented with smaller test sets (usually around 24-36 items). Participants' rule use and knowledge level is then assessed based on this sample of items from the problem space. The goal in Chapter 4 is to "map out" the entire problem space for human performance as well.

CHAPTER 4

TESTING THE IMPLICATIONS OF THE NETWORK AND PROBLEM SPACE ANALYSES

Previous research on the balance-scale task has used Siegler's (1976) method of rule assessment to evaluate the performance of both human participants and computer models. In Chapter 3 a different approach was used with neural networks. A neural network was trained to predict balance-scale problems using the entire population of problems and the converged network was analyzed using a number of interpretive approaches. Similarly, the present research with human participants represents a departure from the standard assessment method. In the first section, the rationale for this departure will be reviewed. In the second section, I will outline predictions for human performance that were derived from previous analyses of neural networks and the problem space. In the third section the details and results of the study with human participants will be described. In the fourth section, the human data will be re-analyzed according to Siegler's criteria to replicate the finding that rule assessment varies with the problems and instances selected.

Rationale for Testing Participants on the Entire Problem Space

There currently are no alternative methods for assessing performance on the balance-scale task. To test the implications of the network interpretation and problem space analysis, an approach analogous to that taken in Chapter 3 was adopted. That is, human participants were tested on the entire set of 625 balance scale problems from the five-peg, five-weight version of the task.

Testing participants on all possible balance-scale problems eliminates the potential confounding of traditionally defined problem-types, torque difference, and match/no-match problems that results when a small sample of problems is selected (see Tables 5, 6, and 7). Arbitrary decisions about which levels of torque difference to sample are avoided. As shown in Appendices D and E, previous researchers used arbitrary levels of torque difference, and different values of torque difference for simple versus conflict problems. Rather than sampling from the problem space, performance for the problem space can be "mapped" with respect to both reaction time and accuracy measures. This approach also allows for comparisons with the neural network results.

Empirical Questions and Predictions for Human Performance

A number of predictions for human performance can be derived from the previous analyses of neural network models and of the problem space. The neural network described in Experiment 1 was interpreted as integrating both the weight and distance dimensions by approximating an additive function. The appropriate comparison, therefore, is with participants who consider both dimensions of the task. The predictions for the psychological data, therefore, concern participants who are at the level described as *Rule 3* by Siegler (1976). That is, the focus is on individuals who do not know the torque rule but realize the importance of both weight and distance for making balance predictions. Based on previous research, it seemed reasonable to use undergraduate students as adults typically do not know the torque rule unless they have been formally taught it (e.g., Aoki, 1991; Chletsos et al., 1989;

Siegler, 1976, 1981). A group of adults is appropriate given the departure of testing participants on 625 trials. Additionally, systematic studies of adult competencies are rare (Shultz et al., 1995).

Torque Difference

The first prediction is that *participants' reaction time and accuracy measures should vary as a function of location in the problem space as indexed by torque difference*. This prediction is based on previous research demonstrating a torque difference effect in human participants (Ferretti & Butterfield, 1986) and neural networks (e.g., Shultz & Schmidt, 1991) including converging evidence from the present network simulations (see Experiments 1 and 2). That is, more sophisticated rule assessments were associated with larger absolute differences in torque (weight x distance) for each arm (see Chapter 2).

Interestingly, the torque difference effect (TDE) has not been replicated in the human literature. Ferretti and Butterfield (1986) were the first to report the TDE. Psychological researchers since then have not attempted to replicate the TDE and there has there been no explicit effort to control for levels of torque difference (e.g., Chletsos et al., 1989; Larivée et al., 1987; Normandeau et al., 1989; Siegler & Chen, 1998).[1] Only modeling researchers have made an effort to test their models for the TDE (e.g., Schmidt & Ling, 1996; Schmidt & Shultz, 1991; Shultz & Schmidt, 1991, Shultz et al., 1995).

[1] In fact, some researchers do not even cite the Ferretti and Butterfield work (e.g., Aoki, 1991; Jansen & van der Maas, 1997; McFadden et al., 1987; van Maanen et al., 1989).

More specifically, differences in accuracy are predicted within instances of the conflict problems. According to Siegler's model, when an individual knows that both dimensions are important but does not know the torque rule, all conflict problems (and all instances of conflict problems) should be treated identically. If torque difference does have an effect it should be especially evident on the conflict problems. Note that a TDE *is* predicted for simple problems. Conflict problems are singled out because they cannot be solved without some type of strategy, guessing, or arithmetic computation, whereas simple problems (balance, weight, distance, and both-greater) can be solved via counting and/or decision trees. If location in the problem space is important then the additional information provided by a large TD should compensate for the "muddling" or "confusion" resulting from the conflict in distance and weight cues. For example, Figure 27 illustrates two conflict-distance problems. In Panel A, the torque difference is equal to 1 and in Panel B it is equal to 10. If TD has an effect, it should be easier to say "tip right" to the scale in Panel B because the left weights are on the right-most peg on that arm and the right weights are on the farthest peg on the right side of the scale.

It is an empirical question, however, at *what particular level* of TD an instance of a simple or conflict problem becomes "easier" to solve. Although there is a restricted range of TD for no-match problems (and small numbers of instances in some cases) accuracy will be determined for each level of TD. This issue has gone unaddressed because as mentioned, little research has been done on the TDE and because Ferretti and Butterfield used dissimilar levels of TD for simple and conflict problems (see Appendix D). By testing subjects on the entire range of problems, it may be possible to determine the threshold of "easy" and "hard" with respect to torque difference effect (i.e., if such a threshold exists). In Chapter 3, I noted that the analysis and predictions outlined in Table 4 were done without taking the torque difference effect into account. For example, for *CD-match* problems if a participant was consistently using either a torque or an additive strategy, most responses would be correct. If a participant was answering in a random way then these items would be correct about a third of the time. If a participant used a predominantly weight-only strategy, performance should be incorrect but in the direction of the greater weight. If there is a TDE, then instances with larger TD should be correct more often than expected by chance alone or by the predictions of a strategy.

Most human studies of the balance-scale task do not include reaction time measures. Based on Siegler's decision tree model, if a single strategy (e.g., muddle through) is applied consistently then there should be no difference in RT for different instances of a problem type. Differences are predicted for each problem type overall. For example, balance, weight and distance problems require *two* decisions each, but both-greater and conflict problems require *three* decisions each (see Figure 2). Presumably, problems requiring two decisions prior to a prediction should take less time than predictions requiring three decisions. If problem instances are treated differently based on characteristics like TD, then this would be reflected in a difference in RT as a function of torque difference.

The Match/No-Match Distinction

The second prediction is that with respect to conflict problems, *participants should be less accurate on no-match problems than on a comparable set of match problems* (i.e., all with TD ≤ 4). As just mentioned, it is an empirical question if there are differences in accuracy within this restricted range of TD. The general pattern of performance predicted for each strategy and for each type of no-match problem in Table 4 does not take TD into account. This prediction is derived from the problem space analysis that showed this subset of problems is unique in that these problems cannot be solved using either an additive or a multiplicative strategy. That is, they can *only* be solved using the torque rule. The results of Experiment 2 were taken as evidence that this subset of problems is different and warrants further investigation with human participants. When these problems (8.3%) were removed from the training set, the network converged reliably in fewer sweeps (approximately 100 versus 4000). The network also converged reliably when one hidden unit was removed.

With respect to reaction time (RT), if the same strategy is applied consistently to all conflict problems then no difference in RT should be observed. If these different instances of conflict problem are approached differently then this would be reflected in a difference in RT.

Intra-Individual Variability versus Consistent Strategy

The previous predictions (and the empirical questions in particular) intersect with recent theorizing about, and approaches to, cognitive development. Siegler

(1996) discussed *universalist* approaches in which the goal was to identify *the* strategy used by participants on a particular task and *comparative* approaches in which the goal was to identify contrasting strategies used by different groups of participants (e.g., younger versus older, expert versus novice). These approaches are contrasted with the *cognitive variability* approach. Rather than viewing variability as a source of error or irritation, it is viewed as a pervasive part of both high- and low-level cognition.[2]

Variability was seen in the neural network simulations (Chapter 3). The network responded differentially (i.e., varying levels of hidden unit activities) to problems depending on their location in the problem space. If one strategy is applied across the problem set by human participants (i.e., as predicted by Siegler's model), certain patterns would be evident. For example, regardless of torque difference, accuracy for conflict problems would be either (a) at chance levels for the "muddling through" strategy, (b) correct with some level of predicted error for the additive or torque strategies, or (c) incorrect consistent with the side with greater weight for the weight-only strategy. If, however, participants adjust their strategy depending on particular instances within the problem space, then support should be found for the previous predictions concerning variability in performance with torque difference and instances of conflict and no-match problem types.

Another way of demonstrating variability in performance is to use Siegler's

[2] How Siegler resolves his current theoretical stance with his previous work on the balance-scale task in which he suggests that only one rule is used per stage will be discussed in more detail in Chapter 5.

rule-assessment method and criteria to show that by manipulating the sample of test items used, performance can be differentially classified as one of Siegler's four rules. That is, the same participant can be classified as using a more or less sophisticated strategy, depending on the sample of test items selected. As discussed in Chapter 2, one of the criticisms of the rule-assessment method was that an individual's classification varied with the items used.

In summary, the main predictions for human performance concern differences in reaction time and accuracy as a function of torque difference and the match/no-match distinction. Differences on these measures can be taken to indicate that participants used variable strategies, whereas similarity in these measures can be taken to indicate the use of a consistent strategy.

Experiment 3: Testing Human Participants on the Entire Problem Space

Method

Design

There are 625 unique combinations of balance-scale problem that can be created with the five-peg, five-weight version of the task. These problems can be classified as one of seven problem types: balance ($n = 25$), weight ($n = 100$), distance ($n = 100$), conflict-weight ($n = 88$), conflict-distance ($n = 88$), conflict-balance ($n = 24$), or both-greater ($n = 200$). Conflict problems can be divided into two main types: match and no-match. There are six types of no-match problem (see Table 3). Moreover, the problems can be classified according to the cluster they fall into. The same pattern was found for several networks such that the same problems fell into clusters of the same size. Approximate numbers are: cluster 1 (*n* = 117), cluster 2 (*n* = 50), cluster 3 (*n* = 125), cluster 4 (*n* = 49), cluster 5 (*n* = 131), cluster 6 (*n* = 32), and cluster 7 (*n* = 121).[3] See Appendix F for the cross-tabulation of problem type and cluster. Given the uneven number of instances per cell when crossing cluster and problem type, five blocks of 125 randomly selected problems were presented to participants.

Procedure

Participants were tested on five blocks of 125 trials. Problems were randomly assigned to each block, and order of presentation within each block was random. Each balance-scale problem was presented on a computer monitor and looked similar to the line drawings used in paper-and-pencil versions of the task (e.g., Chletsos et al., 1989; Ferretti et al., 1985; Siegler, 1981). A demonstration program was used to show the participant the "line drawings" of the balance scale. The demonstration ended with four randomly selected problems so that participants were comfortable with the location and configuration of the buttons used to make a response. Participants were instructed to press the button as soon as they had decided on a prediction. Reaction time was measured using the real-time clock of the computer (see Appendix G for the relevant code).

Each trial was initiated by the participant. The participant pressed one of three keys that corresponded spatially to the drawing of the balance task. Participants used the numeric keypad on the right-hand side of the standard keyboard. Specifically, the "1" key represented "left side down," the "2" key represented "balance," and the "3"

[3] Cluster numbers are arbitrary; these are the cluster numbers from Network 4120.

key represented “right side down.” A block of trials took approximately 10 to 12 minutes to complete. After completing the task, participants were asked to answer three questions about the task (see Appendix H).

Participants

Eight subjects were paid an honorarium to participate in the study. Participants were screened for familiarity with the task. All participants were right handed.

Results

The results will be presented in three sections. First, results pertaining to predictions about the torque difference effect will be presented (holding problem type constant). In the second section, I will outline the results of comparisons between match and no-match problems (holding TD constant). In the third sections, the interaction of torque difference and the various classifications of problem type will be explored.

The Torque Difference Effect: Accuracy

Results bearing on the torque difference effect will be presented collapsed across all problem types. When predicting accuracy from torque difference using linear regression, a TDE was evident, $R^2 = 0.62$ ($F = 40.3$, $df = 1, 23$, $MSe = 1620.9$, $p < .001$). Figure 28 shows that there is a strong relationship between accuracy and torque difference but it does not appear to be strictly linear.

The nonlinear regression module in SYSTAT was used to fit a nonlinear function to the data, predicting percent correct from absolute torque difference. Because the data appeared as if it could be characterized by an exponential function, and because percent correct only ranged from approximately 65 to 100, the following equation was provided to SYSTAT:

(4.1) percent correct = 100 - *P1* * EXP(-1 * *P2* * ABS(TD))

In this equation, *P1* and *P2* are two constants whose values can be manipulated to improve the fit between the equation and the data. SYSTAT's nonlinear regression module used the *Gauss-Newton* method to search the space defined by these two parameters to find their optimal values. The optimal values for *P1* and *P2* were those that provided the best fit between the function and the data, where "best fit" was operationalized as the values that provided the smallest sum of squared differences between the two. After twenty-one iterations of searching, SYSTAT determined that *P1* should be equal to 37.660, and *P2* should be equal to 0.251. In other words, the nonlinear equation relating TD to accuracy was:

(4.2) percent correct = 100 - 37.660 * EXP(-1 * 0.251 * ABS(TD)).

The raw R^2 for this equation was 1.00. The *corrected* R^2 was 0.98, as was the *observed versus predicted* R^2.

Other equations (e.g., a logarithmic function) could be fit to the accuracy data. No strong claims are being made about the precise nature of the equation or of the parameter values in the equation. The non-linear equation is used to demonstrate that one variable (i.e., torque difference) can account for almost all of the variance in the accuracy data for all problems.[4]

[4] Recall that there is 0.95 correlation between the torque and the additive equations (see Figure 23). Using the same procedure, the additive equation (i.e., $|[RW + RD] - [LW + LD]|$) accounts for as much variance in accuracy as the torque equation (corrected $R^2 = 0.99$).

In summary, the torque difference effect has been replicated for the entire problem space for the accuracy data. Moreover, the TDE can be described with respect to a single function.

Reaction Time And The Problem Space

In the previous section, an extremely strong relationship was indicated between one (rough) measure of position in the problem space--torque difference, and one measure of human performance--accuracy. In this section I consider whether position in the problem space can be related to another measure of human performance, namely, reaction time. Reaction time is plotted as a function of absolute torque difference in Figure 29. Again, a trend is evident such that reaction time decreases as torque difference increases. Although this trend is clearly more linear than the one that was observed for accuracy data (see Figure 28), linear regression indicated that TD alone did not account for much of the variance in reaction time ($R^2 = .09$; $F = 494.2$, $MSe = 797.1$, $p < .001$).

This poor fit of the linear regression data, however, does not by itself indicate that a relationship between response latencies and the problem space does not exist. While absolute torque difference is related to the position of a problem in the full problem space (see Figure 19), it is a fairly rough estimate of the location of a problem. Perhaps the relationship between response times and the problem space requires a more accurate measure of a problem's position in the space.

When reaction time was plotted in the same coordinates as the two-dimensional problem space (i.e., as a function of *right torque* [*RW* x *RD*] and *left torque* [*LW* x *LD*]), an interesting interactive relationship does appear to emerge. For example, Figures 30 and 31 plot the reaction times in this space for two of the subjects.[5] In general, the slowest reaction times are along diagonal of these plots, indicating that RT is a function of the interaction between right torque and left torque.

To test the existence of this interactive relationship statistically, linear regression was used to predict participants' RT. Two of the independent variables used to make this prediction were right torque and left torque. Because the preceding figures indicate that these two variables interact, the interaction between these two variables (i.e., right torque x left torque) was also used as a predictor. Finally, because there were considerable inter-subject differences in RT,[6] and because the data was obtained from a repeated measures design, subject identity was used as a predictor, as were the interactions between subject and right torque, left torque, and right torque x left torque. This regression equation was highly significant ($F = 974.8$, $df = 7, 4992$, $p < .001$), and accounted for 58% of the variance in reaction time. In other words, a regression equation that used position in the problem space as predictors, and which took into account between-subject variability in response latency, provided an excellent prediction of response times.

In summary, there is a relationship between the problem space and human performance. There were two main measures of performance. The first measure,

[5] Note that when a composite of all subjects is plotted (5000 data points) the regions associated with different reaction times are not distinguishable.

[6] Mean reaction times for participants ranged from 0.730 to 2.185 seconds ($F = 78.51$, $df = 7, 4992$; $MSe = 125.5$, $p < .001$).

accuracy (i.e., percent correct), was less variable measure between-subjects and it was possible to account for that variability with a single non-linear equation. ***Reaction time*** was more variable between-subjects. A substantial amount of the variability in RT could be predicted by using the location in the problem space and individual between-subject differences.

Match versus No-Match Problems

In Chapter 3, the analysis of the problem space revealed that only a small number of problems could not be solved using a general additive strategy. This unique characteristic led to further exploration with neural nets. The simulations in Experiment 2 demonstrated that the removal of these items from the training set resulted in faster convergence. Therefore, it was predicted that accuracy should be better for match than for no-match problems.

To test this prediction with the data for human participants, match problems with a torque difference of four or less were selected because the range of TD for no-match problems is 0-4. There are 80 conflict problems with TD $\leq$ 4 that can be compared with the 52 no-match problems. Within the conflict-balance (CB) category, there are only four no-match versus 20 match problems, therefore, the analyses were done with and without CB problems.

Accuracy. The accuracy data for each type of match and no-match problem are shown for each participant in Table 9 (the mean for each problem type appears in the last row; match problems are shown in columns 4 and 9). A dependent sample t-test was performed on percent correct for match and no-match (collapsed across subtypes). Overall, participants were more accurate on match problems (66.9%) than on no-match problems (34.4%) and this difference was significant (t = 4.66, df = 7, p = .002). When CB problems were removed, accuracy for match versus no-match problems (75.2% and 35.7%, respectively) was also significantly different (t = 5.87, df = 7, p = .001).

Reaction time. Reaction times were predicted to be greater for no-match problems if participants use different strategies depending on problem characteristics. If a single strategy is applied consistently, however, no difference in reaction time should be observed. Participants were faster to respond to the match than the no-match problems. The mean RTs for match and no-match problems were 2.38 (sd = 1.77) and 2.62 (sd = 1.84) seconds, respectively. Median RT for match and no-match were 1.84 and 2.18, respectively. A paired-sample t-test on median reaction times indicated a significant difference (t = -2.27, df = 7, p < .05).

When CB problems were removed, the mean RTs for match and no-match were 2.21 (sd = 1.60) and 2.65 (sd = 1.85), respectively. Median RT for match and mo-match were 1.74 and 2.20, respectively. The difference in median RT was significant (t = -2.89, df = 7, p < .03).

The results of the comparison between match and no-match problems for human participants indicated that these problems were treated differently as indicated by differences in both response latency and accuracy. The prediction derived from the analysis of the neural network in Experiment 2 and the analysis of the problems space was bourne out in the human data.

The Interaction of Torque Difference and Problem Types

Although a TDE was evident for all problems (for both accuracy and reaction time), it is necessary to determine if the TDE holds for both the traditionally defined problems (i.e., as outlined by Siegler, 1976) and the distinction between conflict problems that can be classified as match versus no-match. Does TD have the same effect on all problem types? To address this question, accuracy for different problem types was plotted as a function of torque difference.

Simple problems. Percent correct for the simple problem types (i.e., simple balance, weight, distance, and both-greater) is plotted as a function of absolute torque difference in Figure 32. Overall, accuracy was high at all levels of torque difference, with simple balance problems having the lowest mean (93%). The mean percent correct for non-balance problems as a function of TD are shown in Table 10.[7] Although accuracy was quite high overall, Figure 33 illustrates that torque difference did have an effect on reaction times. A regression analysis analogous to that described for all problems was conducted. Left torque, right torque, subject identity, and all interactions were entered into a regression equation to predict RT. These factors all contributed significantly and accounted for 64% of the variance in RT.

Conflict problems. Accuracy (i.e., percent correct) for all conflict problems is shown as a function of torque difference in Figure 34. Again, a non-linear relationship is evident. Using the procedures described for the analysis of all problems, the

[7] Most of these are likely to be "true" errors as opposed to systematic ones. All participants reported making mistakes by hitting a key either too quickly or inconsistent with the prediction they wanted to make.

equation of the line was determined in 14 iterations:

(4.3) percent correct = 100 - 61.930 * EXP(-1 * 0.199 * ABS(TD)).

The corrected R^2 was 0.96 (the raw R^2 was 0.998). Again, torque difference accounts for almost all of the variability in accuracy for conflict problems.

The same linear regression procedure was used to analyze the reaction time data (see Figure 35). Left torque, right torque, subject identity, and all interactions were significant in predicting RT and accounted for 63% of the variance in reaction time.

Match versus no-match conflict problems. The effect of torque difference was evident when the overall means for all conflict problems were plotted (see Figure 34). Recall that conflict problems with a torque difference between 0 and 4 include both *match* and *no-match* problems. In Figure 36, accuracy (percent correct) is plotted as a function of TD for both match and no-match problems. When compared to Figure 34, it is clear that some of the TDE for conflict problems can be accounted for by the no-match problems. When a line is fit to the match problem only, torque difference accounts for 79% of the variability in percent correct.

There is some variability within the different types of no-match types, as shown in Table 11. The last four columns of Table 11 include the percent correct for match and no-match, separated into conflict-weight and conflict-distance problems. In general, accuracy was higher for CW problems than it was for CD problems. This pattern is consistent with differentially relying on the weight cue when uncertain.

Reaction times for match and no-match problems are shown as a function of TD in Figures 37 and 38, respectively, for both *correct* and *incorrect* responses. An interesting pattern emerges for *match* problems in that quicker decisions were more likely to be correct, especially at the lower levels of TD (see Figure 37). When participants took extra time to make a response, the prediction was more likely to be incorrect. A deviation from this pattern occurred for conflict-balance problems (TD = 0). Here, a fast response was more likely to be incorrect. The additional time to respond to these problems resulted in higher accuracy. Recall that there is nothing to distinguish any of the conflict problems, they are labeled for their outcomes alone. This same deviation is shown in Figure 38 for *no-match* problems. For CB problems, longer RTs were associated with correct responses. Unlike match problems though, differences in RT for correct and incorrect responses were not evident for non-balance problems.

Discussion

The previous analyses can be used to support the idea that location in the problem space is an important factor for predicting the amount of time to make a prediction and the accuracy of that prediction. With respect to the task *as a whole*, either torque or additive difference (used as an index of location in the problem space) accounts for almost all the variability in accuracy (see Equation 4.2 and Footnote 4). When the problem space was analyzed by different types of problems, the torque difference effect (TDE) was evident in either reaction time or accuracy measures. For example, accuracy was high for simple problem types (93% and higher), but the TDE was reflected in participants' reaction times.

As predicted, the accuracy for no-match problems was lower than for a comparable set of match problems. Differences in reaction time were also found for match versus no-match problems. This finding is suggestive of the idea that although these two types of problem are similar superficially, they were approached differently by participants. That is, given two problems with a conflict in the weight and distance cues with a same or similar level of torque/additive difference, why should one take longer to make a prediction than the other? Perhaps when comparing the two sides of a no-match problem, there is a conflicting reaction such that a perceptual judgment is different from an arithmetic judgment (i.e., comparing the differences in left and right weight and left and right distance). This hypothesis is potentially useful in that reaction time differences were seen such that quicker predictions were more likely to be correct (with the exception of conflict balance problems). The reaction time for no-match problems was significantly slower than for match problems, so the slower reaction for no-match was also associated with reduced accuracy.

In the next section, I will examine the results in terms of Siegler's rule assessment method by showing accuracy data for the traditional problem types and how subjects would be classified if Siegler's method were used.

Reanalysis Using Siegler's Rule Assessment Method

Although Siegler's rule assessment method has been subject to critical analysis, participants' performance will be assessed using Siegler's scoring criteria. The main rationale was to provide further evidence that this assessment technique results in different classifications depending on the instances selected for the test set. As mentioned, a criticism of rule assessment is that classification of performance varies with TD levels (see Chapter 2), but no replication of this variability has been reported since Ferretti and Butterfield (1986). The criticism that assessment can vary with the priority given to the scoring of the various rules has not been addressed in the human literature (Shultz et al., 1994). Both of these issues will be addressed.

Results and Discussion

The accuracy data for each subject by the traditional problem types is shown in Table 12. The general pattern fits with that predicted for Siegler's *Rule 3* (cf. Table 1). Accuracy was very high for balance, weight, and distance problems (columns 3-5). Accuracy was lower for conflict problems (columns 6-8). For most participants though, accuracy on the conflict problems was above chance level but still not high enough to be consistent with the pattern predicted for Rule 4 (i.e., perfect performance for all problem types; see Table 1). These participants were not aware of the torque rule (see Appendix H). Participant 3 indicated a familiarity with the concept of torque but could not define it. It is unclear from the explanation whether she did know the torque rule, as her strategy was illustrated with respect to a simple distance example. Participant 8 indicated a familiarity with torque, but described a non-technical connotation (i.e., "to fine tune something").

Classification for Different Test Sets

Rule classifications were done according to the criteria outlined by Siegler (1981). Participants' accuracy data for the specific problems from previously published test sets were singled out for this assessment. Rule diagnosis was automated using the same program Schmidt and his colleagues used for assessing the performance of computer models (e.g., Schmidt & Ling, 1996; Schmidt & Shultz, 1991, 1992; Shultz & Schmidt, 1991).[8]

Table 13 shows the rule classifications for each subject for a number of different test sets. Five of the test sets were discussed in Chapter 3, including McClelland's test set (adapted from Siegler) and the four test sets used by Shultz and Schmidt (1991) to test connectionist models for the effects of torque difference. Ironically, if comparing performance across these four test sets was the only method of evaluation, a clear effect of TD would not have been found for human participants (see Table 13, columns 3 through 6). To be consistent with the torque difference effect, higher rule classifications should be associated with higher levels of torque difference. The performance of three participants was classified as the same rule for all four levels of TD (i.e., participants 1, 4, and 7), three were classified at one stage higher at either the third or four level of TD (i.e., participants 2, 3, and 8), and two had variable classifications (i.e., participants 5 and 6).

One additional test included only *match* problems for the conflict items (Table 13, column 7), and another test included *no-match* problems for the conflict items (column 8). The results are consistent with the previous analyses that demonstrated that participants were more accurate on match problems. Six of the participants received a lower rule classification on the set with no-match items when compared to the set with match problems. One participants received a higher rule classification, and one received the same classification.

[8] Thanks to William Schmidt for supplying his rule diagnosis code.

The last set included problems with large torque differences (Table 13, column 9). All but one participant was classified at the Rule 4 level for this test set. Participant 7 did not get any conflict-balance problems correct and therefore could not be classified as Rule 4. As mentioned, the written protocols suggest that this group of participants was unfamiliar with the torque rule and yet the response profile can be manipulated to be consistent with Rule 4 by selecting particular instances.

It would be possible to create a large number of sets using either a random selection or selecting systematically within various sets of constraints (e.g., torque difference levels, match versus no-match). The eight test sets in Table 13, however, are sufficient to demonstrate that a consistent strategy was not used for all problems.

Classification for Different Scoring Priorities

Shultz and his colleagues (e.g., Shultz & Schmidt, 1991; Shultz et al., 1994) were the first to observe that a pattern of performance can be consistent with both Rule 2 and Rule 3. Because of this, Shultz and Schmidt (1991) gave priority to Rule 2 when the criteria for both rules were satisfied because fewer errors are predicted for conflict-weight problems (see Table 1). That is, Rule 2 was given priority because systematic, correct performance is predicted for CW problems (versus the chance level performance predicted for Rule 3). Shultz et al. (1994) noted that the issue of assessment order has not been acknowledged or studied in the psychological literature thus far.

Three different *assessment orders* are shown in Table 14. One could determine, for example, if performance is consistent with Rule 4, then Rule 3, then Rule 2, and then Rule 1 (other orders tested were 1, 2, 3, 4 and 4, 2, 3, 1). These three orders were used for three different test sets to evaluate the human data. In columns 2-4 are the results for the test set used by McClelland (1989). Two participants received different rule classifications depending on the scoring priority (i.e., participants 1 and 7). When the scoring priorities were applied to the test set that included only *no-match* conflict problems (columns 5-7), three participants received variable classifications (i.e., participants 1, 2, and 6). Six participants received variable classifications when the test set with only *match* conflict problems was used (columns 8-10). Clearly, scoring priority does affect rule classification when evaluating adult participants. When examining the two assessment orders that could be selected on *a priori* grounds (i.e., 4,3,2,1 and 4,2,3,1), there were only three participants (2, 6, and 7) who received different classifications depending on the assessment order.

Chapter Summary

The experiments in the present chapter were motivated by the analysis and interpretation of artificial neural networks presented in Chapter 3. Specifically, the cluster analysis performed on the hidden units of a converged network resulted in an interesting carving of the problem space. Moreover, there was evidence that the network might be approximating an additive function. The analysis of the problem space revealed that the majority of problems could be solved using an additive heuristic, and that only a small part of the problem space required the use of the torque algorithm for a correct predictions (i.e., no-match problems). The experiments were also motivated by the review of the literature in Chapter 2 in which many criticisms of Siegler's rule-assessment method were brought to light.

One of the criticisms was that rule classification varied with the torque difference of the instances used in the test set (Ferretti & Butterfield, 1986). The torque difference effect, thus far, has not been replicated with human participants. The present results support the existence of the TDE, but interestingly, if Siegler's rule-assessment method was the only measure of performance the results would not clearly support the existence of a TDE. Shultz and Schmidt's (1991) test sets for four levels of TD did not provide unequivocal evidence for a torque difference effect.

An issue that has gone unaddressed in the psychological literature is the discovery made by Shultz and his colleagues concerning the variability in assessment that resulted when using different scoring priorities to evaluate neural networks (see also Schmidt & Ling, 1996). Variable classifications occurred when different scoring priorities were used to evaluate participants' performance and these variable classifications were different for each of the three different test sets used. Many Rule 4 classifications resulted even though none of the participants reported an explicit knowledge of the torque rule.

The overall results converge on the idea that "strategy" varies with location in the problem space. There was no evidence in either the primary analyses or the rule-assessment analyses to support the idea that any participant used *one* strategy consistently for all problems and all instances. This issue of variability of performance, and Siegler's (1996) recent views on the this issue will be explored in Chapter 5.

CHAPTER 5

GENERAL DISCUSSION

In the first section of this chapter, I will summarize the main findings from the empirical studies in Chapters 3 and 4. In the second section I will discuss the implications of the present work for the rule-assessment method. The third section includes implications of this type of approach (i.e., analyzing neural networks and the problem space) for future developmental studies. In the fourth section I will discuss the implications that new conceptions of development (i.e., variability in strategies versus stage-like progressions of single strategies) have for future studies of modeling cognitive development.

Summary of Research

Neural networks were trained to predict balance-scale problems and one of the converged networks was analyzed in detail. Several approaches were used to interpret the structure of the converged network, beginning with an analysis of the connection weights. This analysis provided the initial insight that the network was able to discriminate the difference between left and right weight and between left and right distance based on the configuration of connection weights between the input units and the hidden units and between the hidden units and the output units. The organization of the connection weights was taken as evidence that the network was able to integrate the weight and distance information.

The hidden unit activities were then analyzed in a number of ways. There were regularities in the hidden units based on continuous aspects of the problem. This led to the hypothesis that the network might be approximating a function rather than classifying patterns. Evidence for function approximation was found in the almost perfect correlation between the net input to the hidden units and the torque rule (i.e., the mathematically correct solution). An even higher coefficient resulted when net input was correlated with an *additive* equation (i.e., ([*RW* + *RD*] - [*LW* + *LD*]).[1]

A cluster analysis of hidden unit activities led to a consideration of the *problem space* and how it was "carved" by the neural network. By considering the task in the context of a problem space, it was revealed that the overwhelming majority of problems (91.7%) could be solved correctly by adding the information on each side of the fulcrum and making a decision based on the larger sum. This insight into the problem space provided support for why the network was likely approximating an additive function. Additional support came from the fact that multiplication is not a primitive operation of the processing units in the neural network.

The uniqueness of the problems in which an incorrect prediction results from an additive strategy (i.e., the *no-match* problems) was established by a network trained without these items in the training set. That is, 8.3% of the problems were removed from the training set--resulting in a total of 573 training problems versus 625. When the same topology was used, over 20 runs the networks converged with approximately 100 sweeps of training versus approximately 4000 when the 52 no-match items were included. The network trained on the 573 *match* problems converged reliably when one hidden unit was removed. An analysis of this network revealed distinctive bands

[1] Note that [*RW* + *RD*] - [*LW* + *LD*] is algebraically equivalent to [*RW* - *LW*] + [*RD* - *LD*].

and microbands of hidden unit activity. These bands correlated with the solution to the additive equation.

An analysis of published test sets showed that other researchers have confounded problems classified as match and no-match with various levels of torque difference. Torque difference and the traditionally defined problem types were often confounded. When torque difference was controlled for, different levels of TD were selected for simple versus conflict problems. The interpretation of previous results and the validity of cross-study comparisons is called into question based on these findings.

Based on the analyses of the neural networks, the problem space, and previously published test sets, a number of predictions for human performance were derived. The key predictions were that performance should vary as a function of location in the problem space (using torque difference as a rough index)[2] and that performance should differ for the *match* versus *no-match* problems. An approach analogous to the network simulations was used when testing human participants. The entire set of 625 problems was used in order to "map out" performance in the problem space.

A torque difference effect (TDE) was evident for performance on the whole task. Torque difference (i.e., | left torque - right torque |) was the best predictor of the accuracy data. Using non-linear regression, TD accounted for almost all of the variance in *percent correct*. With respect to *reaction time*, a TDE was evident when

[2] Additive difference could also have been used given the 0.95 correlation between additive difference and torque difference.

using a different index of location in the problem space (i.e., left torque, right torque, and their interaction) and when taking inter-subject variability (plus interactions) into account. Torque difference effects occurred for both conflict problems and simple problems. For simple problems, accuracy was high for all levels of TD but the torque difference effect was observed in the reaction time data.

When performance was reevaluated using the rule-assessment approach and the items used in previously published test sets, the torque difference effect was *not* found. This finding may explain why replications of the TDE have not been reported since Ferretti and Butterfield (1986) first indicated that variable classifications result depending on the TD of the items used. Moreover, sampling various instances from the problem space within other constraints (e.g., all conflict problems were either match or no-match, high levels of torque difference) and evaluating performance resulted in variable classifications when the rule-assessment method was used. Variable classifications were also found when different scoring priorities (i.e., assessment orders) were applied to three different test sets. This issue has not been addressed previously in the psychological literature (Shultz et al., 1994).

Implications for Rule Assessment

The present research converges on the idea that there is intra-individual variability in performance depending on where a problem is located in the problem space. This idea has implications for Siegler's (1976, 1981) account of the balance-scale task and for his rule-assessment method. Recall that in Siegler's earlier work, he suggested that "the basic assumption underlying the rule-assessment approach is that cognitive development can be characterized in large part as the acquisition of increasingly powerful rules for solving problems" (Siegler, 1981, p. 3).

More recently, Siegler (1996) admitted that a "nagging question" arose with respect to "how broadly such rule sequences applied" (p. 58). He identified two distinguishing characteristics of the balance-scale task and other developmental tasks (e.g., the projection of shadows task, conservation of liquid, the slopes task) in which sequences of rules have been observed. First, they are *unfamiliar* tasks that children and adults typically do not encounter outside a laboratory setting. Second, they have two or more discrete features or dimensions, one of which is usually more salient for children and dominates their judgments.

Siegler (1996) argued that these two characteristics could be responsible for the "especially frequent use of a *single, consistent strategy*" (p. 59; emphasis added). As mentioned in Chapter 4, Siegler (1996) is now a proponent of the *cognitive variability* approach to cognitive development. This approach contrasts with approaches designed to identify *the* strategy used by participants on a particular task (i.e., the *universalist* approach) or to identify the strategy used by different groups of participants (e.g., younger versus older) on a particular task (i.e., the *comparative* approach). In the cognitive variability approach, the basic assumption is that "at any one time, children have a variety of ways of thinking about most topics" (Siegler, 1998, p. 92).

Siegler (1996) reconciles his current theoretical stance with previous work via the *moderate experience hypothesis.* That is, Siegler has suggested that *multiple* strategies are most likely to be available when an individual has a moderate amount of experience with a particular task. Figure 39 illustrates this hypothesis as an inverted U-shaped curve. On the abscissa is *prior experience*, shown as a continuum of low, moderate, and high. On the ordinate is the number of strategies available, also shown as a continuum. When an individual first encounters a problem solving situation, they have few strategies at their disposal. Similarly, once an individual has gained some level of expertise, they know and use a small number of efficient strategies. It is when an individual has *moderate* experience with a problem solving situation that a variety of strategies are likely to be observed.

So the inconsistency in Siegler's balance-scale work and his more recent views on the importance of variability in strategy use is resolved by appealing to peculiar characteristics of the balance-scale task. That is, a single consistent rule can be applied in the case of the balance-scale task because it is unfamiliar and has a dominant and a subordinate dimension. The cumulative evidence based on the review of the literature in Chapter 2 and the present research, however, is suggestive of variability in strategies on the balance-scale task and raises the question of whether any reconciliation is really necessary. Consider the list of criticisms of rule assessment summarized in Chapter 2: (a) assessment varying with task demands, (b) assessment varying with the torque difference of the items used, (c) assessment varying with the priority given to the various rules, (d) assessment varying with scoring criteria, (e) scoring criteria that are not diagnostic with respect to other postulated rules, and (f) lack of clarity regarding the "muddle through" stage. The first four criticisms have to do with *variability* in assessment, because the goal of the rule-assessment method has been to assign *one* rule per individual and per stage of development.

The criticisms listed above were described as having to do with either the *underestimation* of children's knowledge about balance concepts or to the *incorrect classification* of participants. Criticisms revolved around the fact that the method was either too conservative and/or too restrictive with respect to falsely classifying an individual's one "true" strategy or knowledge level. Given the variability in performance measures in the present study (i.e., accuracy and reaction time), it is unlikely that a single category label (e.g., "Rule 3," "Additive rule," "Qualitative proportionality") would adequately describe any particular participant's performance or strategy, especially as one that was applied consistently, across all instance of balance-scale problem.

In summary, there is no need to appeal to task characteristics or the moderate experience hypothesis to achieve consistency in older and newer ways of evaluating performance. Clearly, the *goal* of rule-assessment has been consistent with the universalist approach--to find *one* invariable or modal strategy for each individual or stage of development. The issue is not with characteristics of the balance-scale task (or isomorphs) but the objectives of the approach. Until recently, variability in rule-assessments was seen as problematic. Where reconciliation *is* necessary is with the goals of rule-assessment and new conceptions of development.

Implications of the Network Interpretation Approach for Future Studies of Development

The balance-scale task has been of interest to developmental psychologists because of age-related trends in performance. Clearly, a required next step is to test some of the implications of this research with participants from different age groups. The current research has been focused on understanding the task itself as a precursor to understanding performance on the task. The empirical predictions derived from studying the task with neural networks and from analyzing the problem space were tested on undergraduate students. This group of participants was at level at which their judgments were not dominated by a single dimension of the task.

One of the issues debated in the literature has been whether connectionist models should capture Rule 4 behavior or not, as it has been asserted that few individuals reach this level of performance without explicit instruction. A prediction derived from the cascade-correlation models described in Shultz et al. (1995) was that Rule 4 behavior may not require an explicit knowledge of the torque rule. Shultz et al. asserted that this level of competence could be achieved by "adequate exposure to the problem domain." Given the present results, it would appear that performance consistent with Rule 4 does not require prior exposure. Given the nature of the task, it is possible to observe such behavior depending on the instances selected for the test set. Moreover, given the analysis of the problem domain, there are few instances in the problem space that can be used to discriminate between individuals using the torque rule or an additive rule (i.e., based on non-verbal patterns of performance).

With respect to deriving predictions for younger participants, the neural network models described in Chapter 3 were integrating weight and distance, and thus may not be an appropriate source of hypotheses. Network simulations trained to perform consistent with Rule 1 and Rule 2 descriptions may provide predictions for human performance. That is, neural networks can be trained to perform the task in a different way (incorrect, but consistent with the performance of children). These networks can then be interpreted, and the problem space explored.

The results of pilot studies of this nature are suggestive of the idea that younger participants may be able to perform the task via discrete pattern recognition. If this is the case, predictions for performance might include an effect analogous to either the torque difference effect or the "problem size effect" in mental arithmetic (e.g., LeFevre, Sadesky, & Bisanz, 1996). That is, accuracy and reaction time should vary as a function of the size of the difference between left weight and right weight. Performance may also vary with respect to the number of weights and pegs, given differences in performance because of subitizing versus counting (Folk, Egeth, & Kwak, 1988; Trick & Pylyshyn, 1993, 1994).

Implications of New Conceptions of Development for Modeling Studies

Given the balance-scale task has been re-cast in the light of recent trends that consider the importance of intra-individual variability rather than qualitative, stage-like changes in behavior (e.g., Siegler, 1996), what are the implications for modeling cognitive development? That is, how are new conceptualizations of development going to affect how modeling of cognitive development proceeds? In the previous section, I discussed ways in which research could be directed from exploring neural networks to deriving and evaluating predictions with children or adults. But consider research focused in the *other* direction, that is, using patterns of human performance (and developmental patterns in particular) and modeling them with connectionist or symbolic models. This line of research typifies previous modeling research on the balance-scale task. One reason that connectionist models of the balance-scale task were considered significant was because they demonstrated that stage-like or qualitative shifts in performance could result from small, continuous, quantitative changes (Bates & Elman, 1992; McCleod et al., 1998; Shultz, 1991).

Now that performance on the balance-scale task may be interpreted as subject to the same variability in strategy use as other, more "everyday" tasks (Siegler, 1996) it remains to be seen how connectionist or symbolic models will need to be modified in order to simulate this type of behavior pattern. Computer models that are consistent with current theorizing about importance on the concept of cognitive variability as opposed to stage-wise depictions of development are currently being developed. For example, Siegler and Shipley's (1995) Adaptive Strategy Choice Model (ASCM) model was developed to account for the selection of strategies on simple arithmetic problems. Klahr and MacWhinney (1998), however, refer to ASCM as an *ad hoc* model. That is, it is described as a type of computer model that "employs an ad hoc computational architecture in which to formulate and run the model" (p. 634) but does not include the systematic assumptions or theoretical commitments of connectionist or production system approaches (Klahr & MacWhinney, 1998). With respect to the connectionist and production system frameworks, new sets of assumptions, and perhaps a new generation of models, will be required to capture the variability in strategy observed on various tasks (e.g., arithmetic, scientific reasoning) in which observed behavior is consistent with the "overlapping waves" depiction of development.

Final Remarks

There has been a growing interest in modeling cognitive development (e.g., Bates & Elman, 1990; Elman, 1993; Elman et al., 1996; Klahr & MacWhinney, 1998; Mareschal & Shultz, 1996; McClelland, 1995; Plunkett et al., 1997; Plunkett & Sinha, 1992; Shultz et al., 1995). Recent characterizations of developmental process have important implications for modeling developmental processes–implications that open up a new set of challenges for researchers interested in modeling cognitive development and evaluating the performance of those models. The enterprise may benefit from an approach similar to that adopted here: one that shifts among the analysis of neural networks, analysis of the task, and analysis of human performance. It is through a reciprocal approach that future researchers may face these new challenges.

Table 1

Predicted Success (Percentage of Correct Responses) on Different Balance-Scale Problems for Individuals Using Siegler's (1976) Four Rules

Problem Type	Level of Performance			
	Rule 1	Rule 2	Rule 3	Rule 4
Balance	100	100	100	100
Weight	100	100	100	100
Distance	0[a]	100	100	100
Conflict-Weight	100	100	33[c]	100
Conflict-Distance	0[b]	0[b]	33[c]	100
Conflict-Balance	0[b]	0[b]	33[c]	100
Both-Greater[d]	100	100	100	100

[a] Incorrectly predict that the scale will balance. [b] Incorrectly predict that the side with the greater weight will go down. [c] Chance responding is predicted. [d] Problem type not discussed by Siegler (1976, 1981). One side has the greater weight and the greater distance.

Table 2

Characteristics of the Clusters for Network 4120

Cluster [a]	n	Mean Hidden Unit Activity				Mean Torque
		H0	H1	H2	H3	Difference [b]
1 (left)	117	.84 (.12)	0	0	0	-11.47 (5.07)
7 (left)	121	.38 (.11)	.08 (.13)	.03 (.02)	0	-5.97 (4.06)
6 (left)	50	.16 (.03)	.80 (.16)	.08 (.03)	.02 (.03)	-2.20 (1.40)
3 (balance)	49	.09 (.01)	.84 (.04)	.14 (.01)	.54 (.03)	0
2 (right)	127	0	0	.87 (.11)	0	10.88 (5.22)
5 (right)	129	.02 (.01)	.02 (.03)	.43 (.12)	.03 (.06)	5.66 (4.17)
4 (right)	32	.06 (.01)	.31 (.09)	.21 (.02)	.71 (.14)	1.94 (1.22)
Total	625	.26 (.32)	.17 (.30)	.30 (.33)	.09 (.21)	0 (8.95)

Note. Standard deviations are shown in brackets. When a mean of zero is shown without brackets, the standard deviation is also zero.

[a] Clusters have been reordered by output state of the network (shown in brackets).

[b] Torque difference = (right weight x right distance) - (left weight x left distance).

Table 3

Characteristics of the 52 Problems Not Solvable by an Additive Heuristic

Problem Type	Torque predicts:	Additive predicts:	Frequency
Conflict Balance	balance	tip (> weight)	2 [a]
Conflict Balance	balance	tip (> distance)	2 [b]
Conflict Weight	tip	balance	20 [c]
Conflict Weight	tip	tip opposite side	4 [d]
Conflict Distance	tip	balance	20 [e]
Conflict Distance	tip	tip opposite side	4 [f]

Note. All absolute torque difference values for conflict-weight and conflict-distance problems range between 1 - 4.

[a] Conflict balance sum weight (CBSW). [b] Conflict balance sum distance (CBSD).

[c] Conflict weight sum balance (CWSB). [d] Conflict weight sum distance (CWSD).

[e] Conflict distance sum balance (CDSB). [f] Conflict distance sum weight (CDSW).

These labels were suggested by Ferretti et al. (1985).

Table 4

Predictions Made by Three Different Strategies for Six Types of No-Match Problems and Three Types of Match Problems

	Strategy Type			
No-Match Type	Torque	Additive	Weight Only	Frequency [a]
CBSW	BAL	>WT	>WT	2
CDSW	>DIS	>WT	>WT	4
CWSB	>WT	BAL	>WT	20
CWSD	>WT	>DIS	>WT	4
CBSD	BAL	>DIS	>WT	2
CDSB	>DIS	BAL	>WT	20
Match Type				
CB	BAL	BAL	>WT	20
CD	>DIS	>DIS	>WT	64
CW	>WT	>WT	>WT	64

Note. CBSD = conflict-balance sum-distance, CBSW = conflict-balance sum-weight CDSB = conflict-distance sum-balance, CDSW = conflict-distance sum-weight CWSB = conflict-weight sum-balance, CWSD = conflict-weight sum-distance. BAL = scale balances, >DIS = tips to side with greater distance, >WT = tips to side with greater weight.

[a] Out of 625, for a five-peg, five-weight version of the task.

Table 5

Number of Problems in the Test Set Used by McClelland (1989) and Schmidt and Shultz (1991) by Torque Difference (TD) and Problem Type

Torque	Problem Type				
Difference	CD	CW	DIS	WT	Total
1	1	0	0	0	1
2	1[a]	1[a]	1	0	3
3	0	1	0	0	1
4	2	1	1	1	5
5	0	1	0	0	1
6	0	0	1	0	1
7	0	0	0	0	0
8	0	0	0	3	3
9	0	0	0	0	0
10	0	0	1	0	1
Total	4	4	4	4	16

Note. Absolute value of torque difference = $|(RW \times RD) - (LW \times LD)|$.

CD = conflict-distance, CW = conflict-weight, DIS = distance, WT = weight.

[a] No-match problem.

Table 6

Number of Problems in the Test Set Used by Jansen and van der Maas (1997) by Torque Difference (TD) and Problem Type

Torque	Problem Type				
Difference	CD [a]	CW [b]	DIS	WT	Total
1	2	3	1	0	6
2	3	2	1	3	9
3	0	0	1	1	2
4	0	0	1	0	1
5	0	0	0	0	0
6	0	0	1	0	1
7	0	0	0	0	0
8	0	0	0	1	1
Total	5	5	5	5	20

Note. Absolute value of torque difference = $|(RW + RD) - (LW + LD)|$.

CD = conflict-distance, CW = conflict-weight, DIS = distance, WT = weight.

[a] Four of the five were no-match problems. [b] Three were no-match problems.

Table 7

Number of Problems in the Test Set Used by Schmidt and Shultz (1991) for Two Levels of Torque Difference (TD) by Problem Type

Torque	Problem Type			
Difference	CD Match	CD No-Match	CW Match	CW No-Match
Level 1	3	1	2	2
Level 2	2	2	4	0

Note. Level 1 problems were TD = 1; level 2 problems were TD = 3.

Table 8

Band Membership for H1 as a Function of the Additive Equation for Network 2929

Additive	Hidden Unit 1 Band[b]					
Difference[a]	A1	A2	B1	B2	C	Total
-8	1	0	0	0	0	1
-7	4	0	0	0	0	4
-6	10	0	0	0	0	10
-5	20	0	0	0	0	20
-4	35	0	0	0	0	35
-3	0	52	0	0	0	52
-2	0	0	0	68	0	68
-1	0	0	0	0	74	74
0	0	0	45	0	0	45
1	0	74	0	0	0	74
2	68	0	0	0	0	68
3	52	0	0	0	0	52
4	35	0	0	0	0	35
5	20	0	0	0	0	20
6	10	0	0	0	0	10
7	4	0	0	0	0	4
8	1	0	0	0	0	1

Note. There are a total of 573 patterns (match problems only).

[a] Additive difference = (*RW* + *RD*) - (*LW* + *LD*).

[b] Bands A1, A2, B1, and B2 refer to the microbands in hidden unit 1 (see Figure 25).

Table 9

Accuracy Data (Percent Correct) for Each Participant for Conflict Problems with Torque Difference ≤ 4 (Match and No-Match Problems)

Participant	Problem Type							
	CBSD[a]	CBSW[b]	CB Match[c]	CDSB[d]	CDSW[e]	CWSB[f]	CWSD[g]	Match[h]
1	100	50	50	65	25	80	50	78.3
2	0	0	55	30	25	40	25	80.0
3	0	0	10	20	0	60	75	81.7
4	0	0	80	15	0	25	25	90.0
5	0	0	20	0	0	45	25	68.3
6	0	50	70	15	25	45	50	46.7
7	0	0	0	55	25	45	25	81.7
8	0	100	50	60	50	0	0	75.0
Mean	12.5	25.0	41.9	32.5	18.8	42.5	34.4	75.2

Note. The frequency of no-match problems (out of 625) is as follows: CBSD (n = 2), CBSW (n = 2), CDSB (n = 20), CDSW (n = 4), CWSB (n = 20), CWSD (n = 4).

[a] Conflict balance sum distance (CBSD). [b] Conflict balance sum weight (CBSW).

[c] Match conflict-balance problems (n = 20).

[d] Conflict distance sum balance (CDSB). [e] Conflict distance sum weight (CDSW).

[f] Conflict weight sum balance (CWSB). [g] Conflict weight sum distance (CWSD).

[h] Match problems with torque difference ≤ 4 (n = 60).

Table 10

Accuracy as a Function of Torque Difference for Simple Problem Types

Torque Difference	Total Correct	Both-Greater	Weight	Distance
1	98.4		100	96.9
2	97.8		99.1	96.4
3	96.2	93.8	95.8	96.9
4	95.8	90.6	96.1	96.9
5	96.6	96.3	98.4	95.3
6	95.3	89.6	98.8	98.8
7	95.8	95.8		
8	97.1	96.4	100	95.3
9	99.4	99.1	100	100
10	99.5	100	100	97.9
11	99.2	99.2		
12	98.3	100	97.9	95.8
13	99.2	99.2		
14	98.4	98.4		
15	100	100	100	100
16	100	100	100	100
17	97.9	97.9		
18	100	100		
19	100	100		
20	100		100	100
21 - 24	100	100		

Note. Percentage correct for simple balance problems was 93%. Empty cells indicate that a problem type does not have instances at that level of TD.

Table 11

Accuracy for Conflict Problems as a Function of Torque Difference Level

Torque Difference	Total Correct	CW Match	CW No-Match	CD Match	CD No-Match
1	54.4	85	33.8	68.8	30
2	54.3	75	42.2	65.6	34.4
3	63.1	81.3	50	79.2	25
4	65.5	87.5	56.3	60.4	25
5	81.9				
6	79.9				
7	89.8				
8	90.6				
10	89.1				
11	95.3				
15	93.8				

Note. The total percent correct for conflict-balance problems was 38.0%. Percent correct for match and no-match was 41.8% and 18.7%, respectively. The range of torque difference for no-match problems is 0-4.

Table 12

Accuracy Data (Percent Correct) for Each Participant as a Function of Problem Type

	Problem Type						
Participant	Both Greater	Balance	Weight	Distance	Conflict-Weight	Conflict-Distance	Conflict-Balance
1	100	88	100	100	86.4	78	54.2
2	96.5	96	99	95	79.5	64.8	45.8
3	98	92	100	98	89.8	61.4	8.3
4	99	100	97	100	77.3	65.9	66.7
5	97.5	92	97	93	73.9	42	16.7
6	93.5	100	97	95	71.6	30.7	62.5
7	100	80	98	97	69.3	80.7	0
8	98.5	96	99	99	50	83	50
total	97.9	93	98.4	97.3	74.7	63.4	38

Table 13

Rule Diagnosis for Each Participant for Eight Different Test Sets

	Test Set							
Ss[a]	McC[b]	S & S level 1[c]	S & S level 2[d]	S & S level 3[e]	S & S level 4[f]	all match[g]	no-match[h]	high TD[i]
1	4	4	4	4	4	2	4	4
2	3	3	3	4	4	4	3	4
3	2	3	3	4	4	4	2	4
4	3	4	4	4	4	4	3	4
5	0	0	2	3	2	2	2	4
6	2	2	3	2	4	4	3	4
7	3	3	3	3	3	3	0	3
8	3	3	3	3	4	4	3	4

Note. Rule diagnosis was performed as outlined by Siegler (1976). A diagnosis of 0 indicates "unclassifiable."

[a] Participant's identification number.

[b] McClelland's (1989) test set published in Shultz and Schmidt (1991).

[c] Shultz & Schmidt (1991) torque difference level 1 (TD = 1).

[d] Shultz & Schmidt (1991) torque difference level 2 (TD = 3).

[e] Shultz & Schmidt (1991) torque difference level 3 (TD = 12 for simple problems; TD = 5 for conflict problems).

[f] Shultz & Schmidt (1991) torque difference level 4 (TD = 15-20 for simple problems; TD = 10-15 for conflict problems).

[g] A sample of problems with all conflict problems classified as *match* problems.

[h] A sample of problems with all conflict problems classified as *no-match* problems.

[i] A sample of problems with high torque difference for non-balance problems.

Table 14

Rule Diagnosis for Each Participant for Three Test Sets and Three Different Assessment Orders

	Test Set								
	McClelland[a]			No Match[b]			All Match[c]		
Ss[d]	4321[e]	1234[f]	4231[g]	4321[e]	1234[f]	4231[g]	4321[e]	1234[f]	4231[g]
1	4	3	4	4	3	4	2	2	2
2	3	3	3	3	2	2	4	3	4
3	2	2	2	2	2	2	4	2	4
4	3	3	3	3	3	3	4	3	4
5	0	0	0	2	2	2	2	2	2
6	2	2	2	3	2	2	4	3	4
7	3	2	2	0	0	0	3	2	2
8	3	3	3	3	3	3	4	3	4

Note. Rule diagnosis was performed as outlined by Siegler (1976). A diagnosis of 0 indicates "unclassifiable."

[a] Test set used by McClelland (1989) and Shultz and Schmidt (1991).

[b] A sample of problems with all conflict problems classified as *no-match* problems.

[c] A sample of problems with all conflict problems classified as *match* problems.

[d] Participant's identification number.

[e] Scoring priority: Rules 4, 3, 2, 1.

[f] Scoring priority: Rules 1, 2, 3, 4.

[g] Scoring priority: Rules 4, 2, 3, 1.

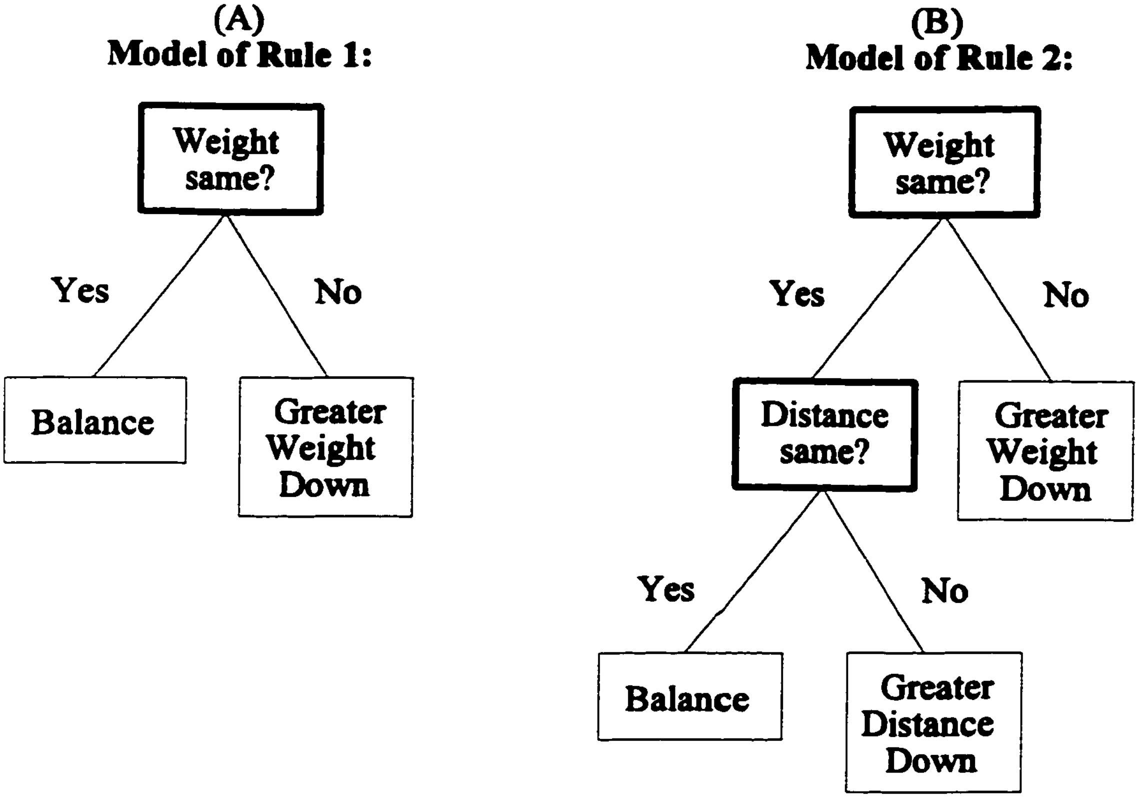

Figure 1. Siegler's (1976) decision tree model (Rules 1 and 2).

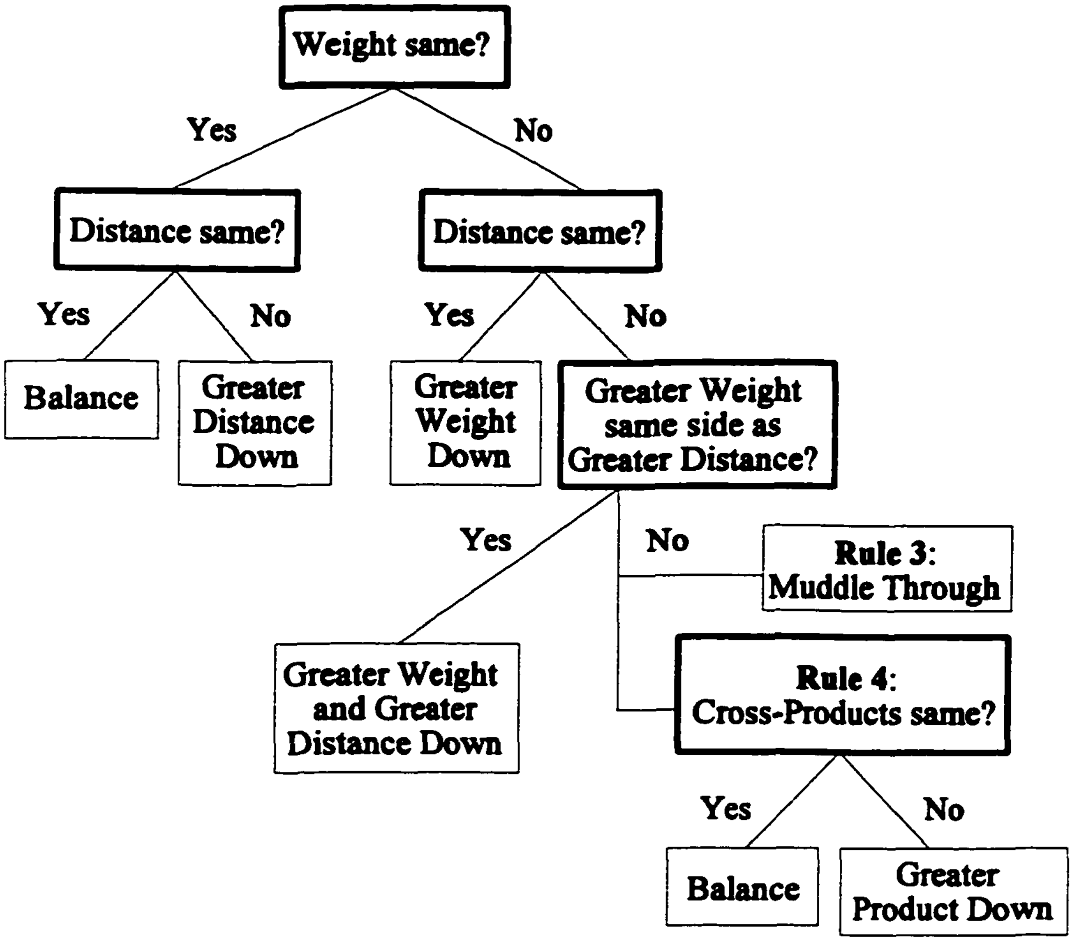

Figure 2. Siegler's (1976) decision tree model (Rules 3 and 4). The difference between Rule 3 and Rule 4 is illustrated in the branch of the decision tree that occurs after ascertaining that the weights and the distances are in conflict.

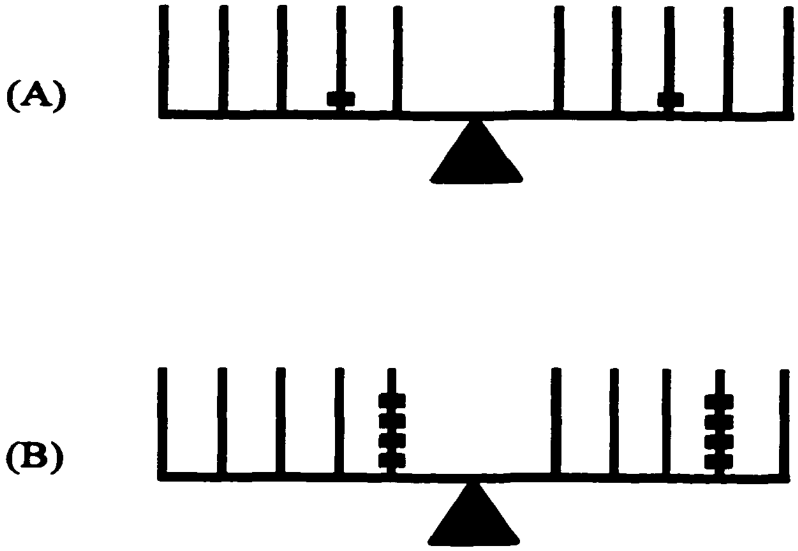

Figure 3. Two simple distance problems used by Ferretti and Butterfield (1986) to illustrate torque difference levels. The torque difference is 1 in Panel A (Level 1) and 12 in Panel B (Level 3).

```
Rule 1
        P1: ((Same W) → (Say "balance"))
        P2: ((Side X more W) → (Say "X down")

Rule 2
        P1: ((Same W) → (Say "balance"))
        P2: ((Side X more W) → (Say "X down")
        P3: ((Same W) (Side X more D) → (Say "X down"))

Rule 3
        P1: ((Same W) → (Say "balance"))
        P2: ((Side X more W) → (Say "X down")
        P3: ((Same W) (Side X more D) → (Say "X down"))
        P4: ((Side X more W) (Side X less D) → muddle through)
        P5: ((Side X more W) (Side X more D) → (Say "X down"))

Rule 4
        P1: ((Same W) → (Say "balance"))
        P2: ((Side X more W) → (Say "X down")
        P3: ((Same W) (Side X more D) → (Say "X down"))
        P4: ((Side X more W) (Side X less D) → (get Torques))
        P5: ((Side X more W) (Side X more D) → (Say "X down"))
        P6: ((Same Torque) → (Say "balance"))
        P7: ((Side X more Torque) → (Say "X down"))
```

Figure 4. The production system code used by Klahr and Siegler (1978). D = distance, W = weight, P = production. One the left side of the "→" is the *condition*, on the right is the *action*.

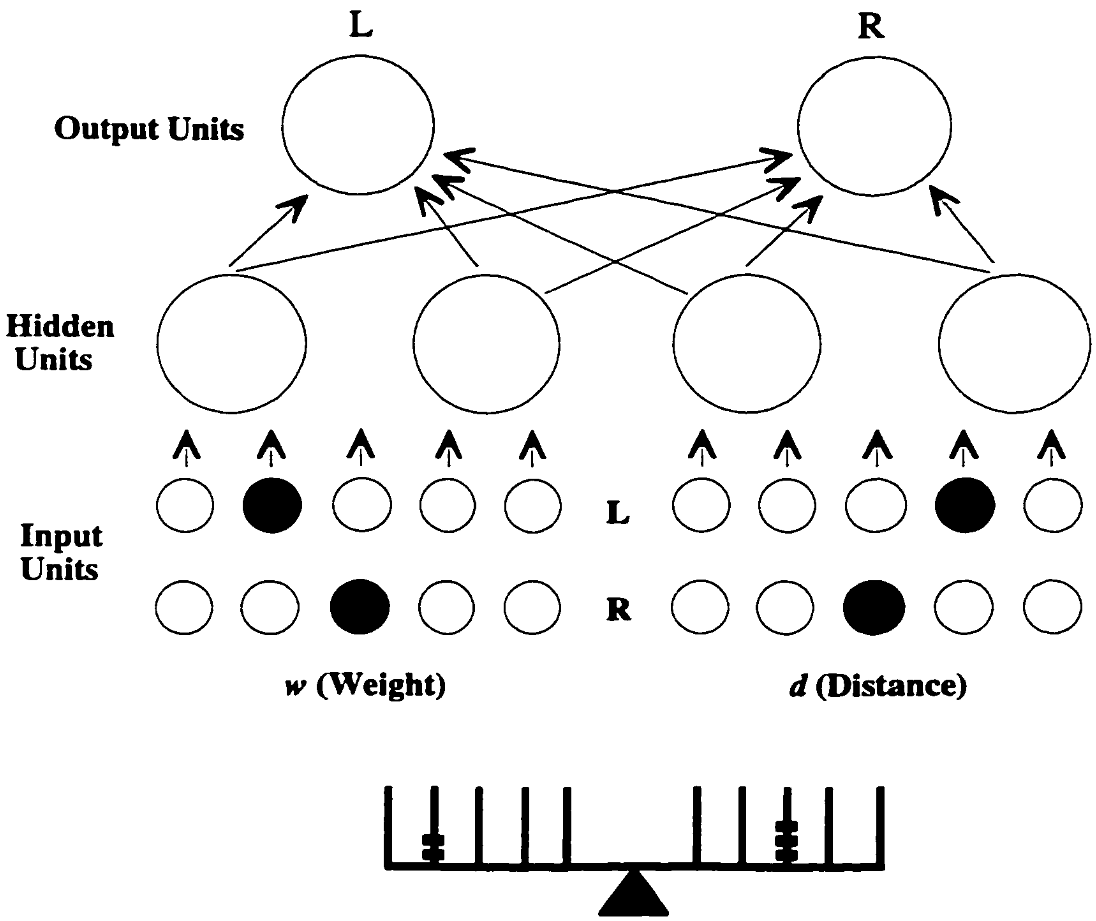

Figure 5. McClelland's (1989) architecture and the balance-scale problem represented on the input units.

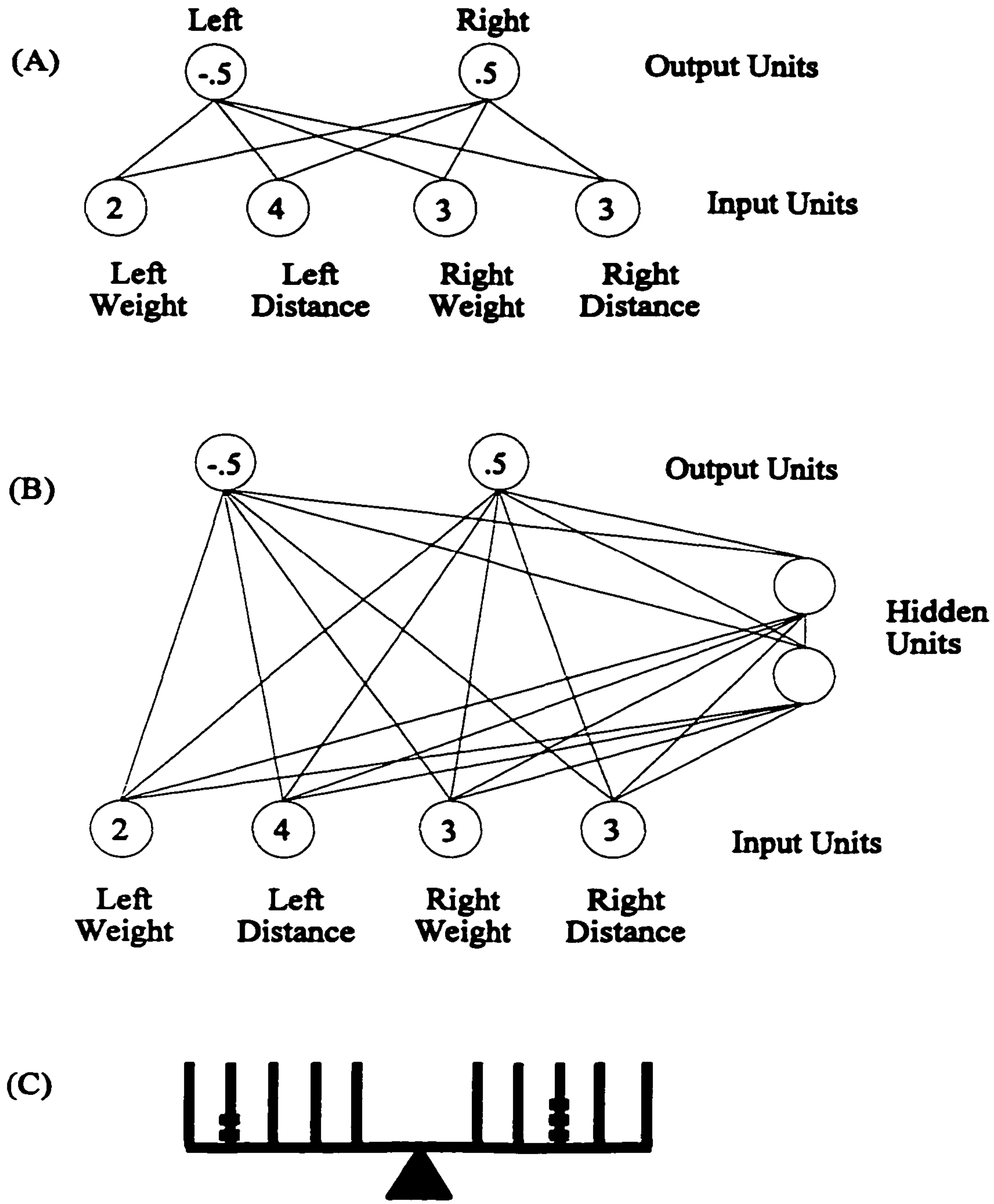

Figure 6. The cascade-correlation architecture used by Shultz and Schmidt (1991). Panel A shows the initial network without any hidden units. Panel B shows the network after the recruitment of two hidden units. The values on the input and output units represent the balance-scale problem illustrated in Panel C. Adapted from Shultz et al. (1994).

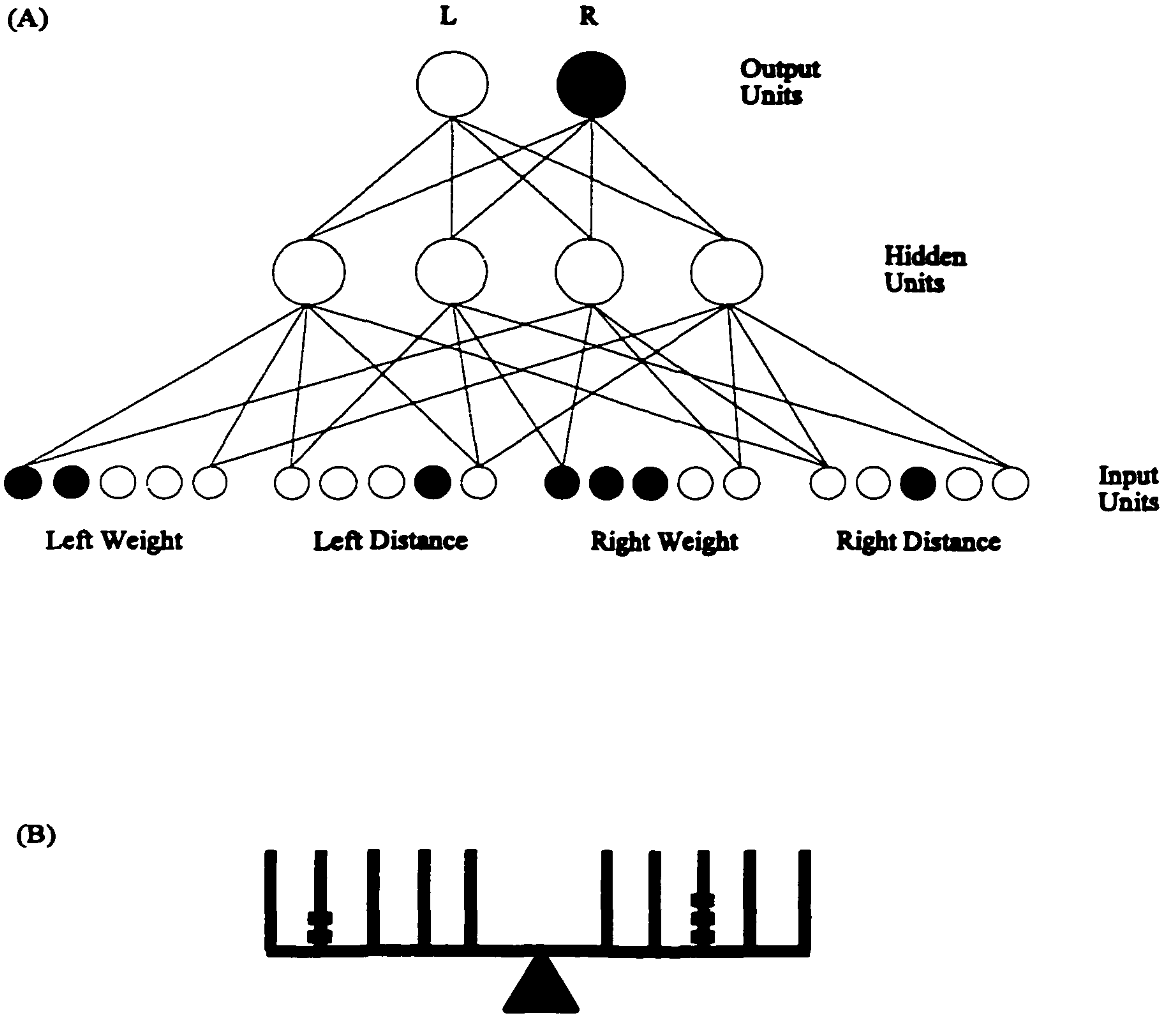

Figure 7. The network architecture used in Experiment 1. Weights are thermometer coded and distances are unary coded. The network is shown with four weights on the fourth peg on the left side and two weights on the third peg on the right side. The network is fully-connected, although not all connections are not shown.

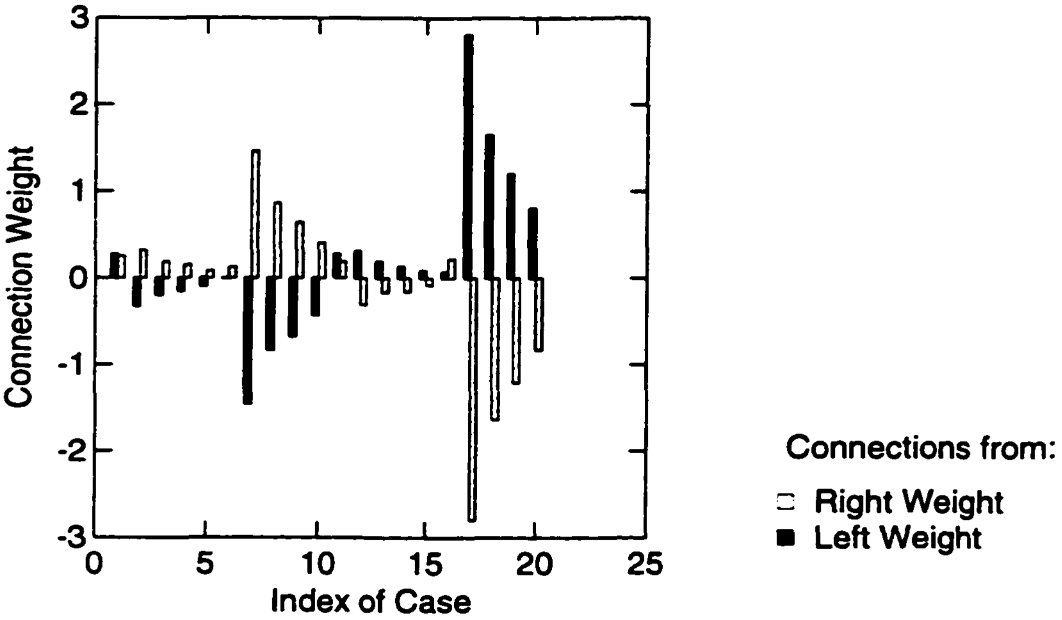

Figure 8. Plot of the connection weights in Network 4120 between the four hidden units and the input units representing left weight (solid bars) and right weight (striped bars) and the four hidden units. The first five bars in each panel are the connections to H0, the second five to H1, the third set of five to H2, and the last five to H3. H1 and H3 are associated with patterns with the smallest torque differences, H0 acts as the left detector and H2 acts as the right detector. See text for further explanation.

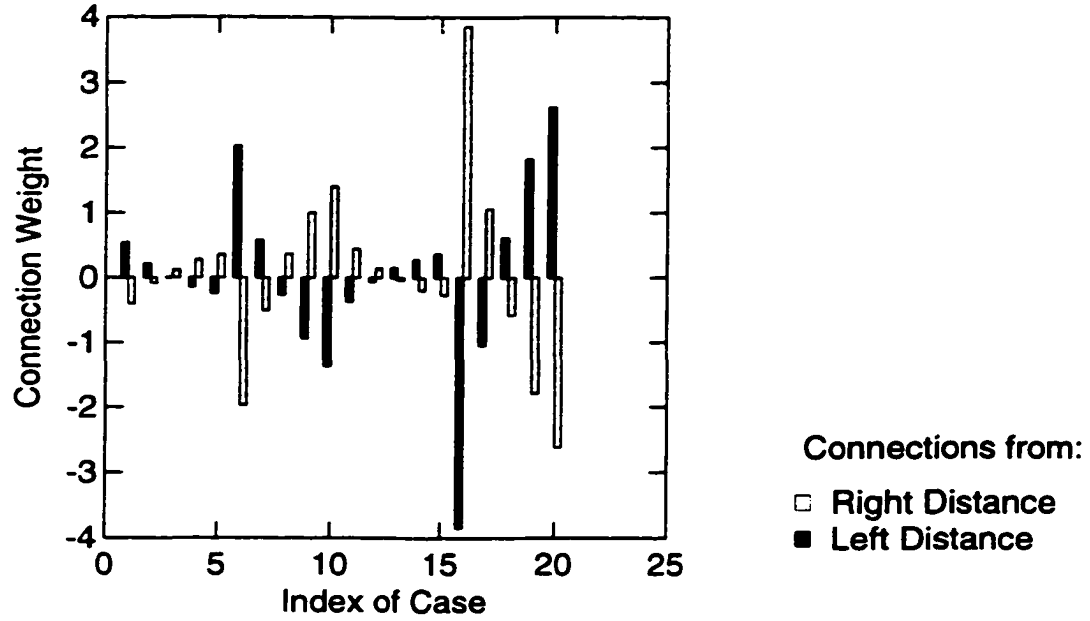

Figure 9. Plot of the connection weights in Network 4120 between the four hidden units and the input units representing left distance (solid bars) and right distance (striped bars) and the four,hidden units. The first five bars in each panel are the connections to H0, the second five to H1, the third set of five to H2, and the last five to H3. H1 and H3 are associated with patterns with the smallest torque differences. See text for further explanation.

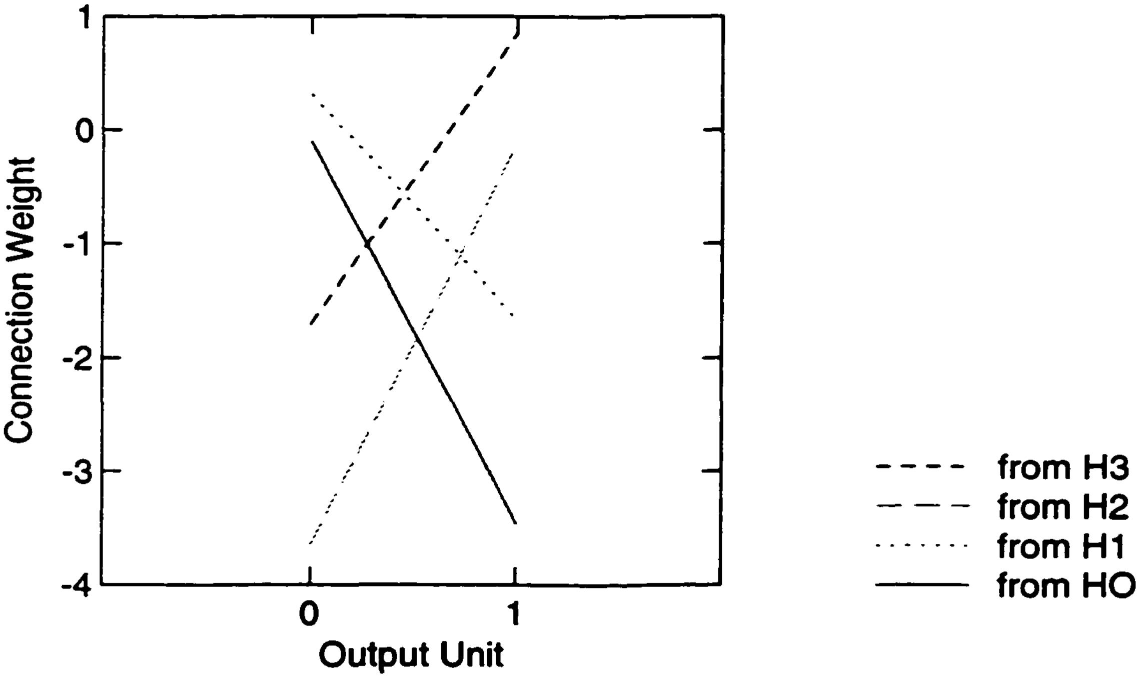

Figure 10. Plot of the connection weights between the four hidden units and the two output units. For any given hidden unit, the connection-weight to one output unit is a large value relative to the value of the connection-weight to the other output unit. See text for further explanation.

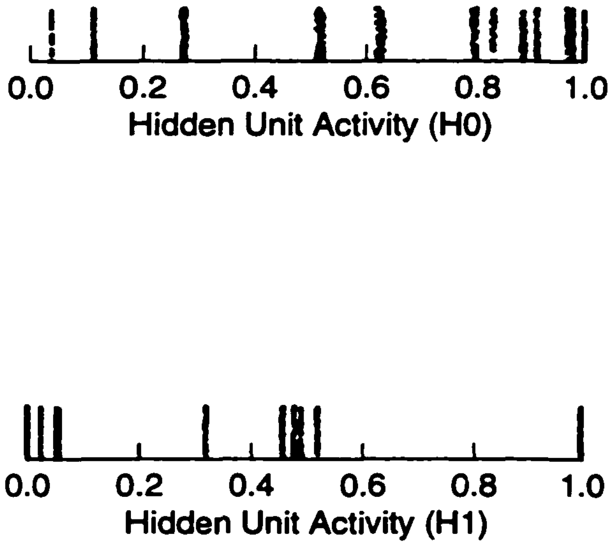

Figure 11. An example of ***banding*** in the hidden unit activities of a converged neural network for purposes of illustration. Typically, each band contains input patterns that have common features that allow the network to learn the input-output mapping.

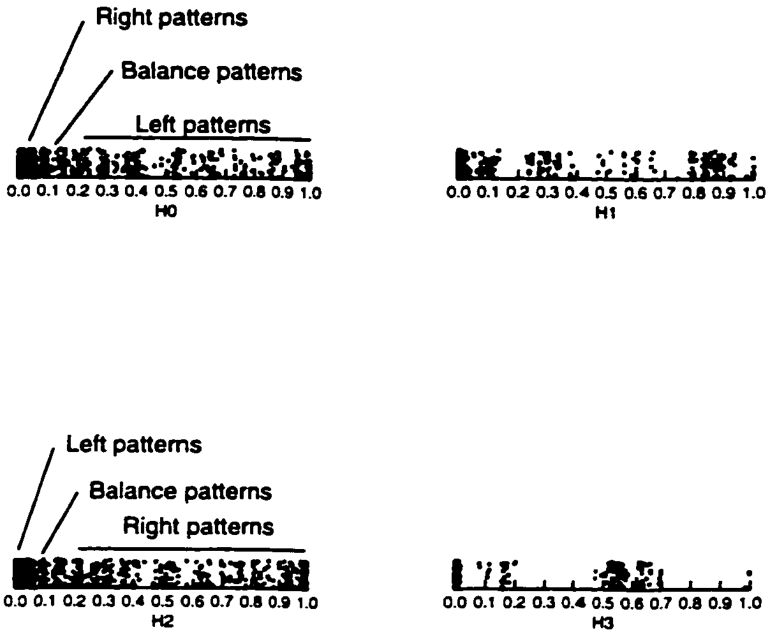

Figure 12. Jittered density plots for the four hidden units in Network 4120. Hidden unit activity is shown on the abscissa. There is a random spread on the ordinate. The approximate regions for regularity in the activities for left, right, and balance problems are shown for two hidden units (H0 and H2). Each dot represents one of the 625 balance-scale problems.

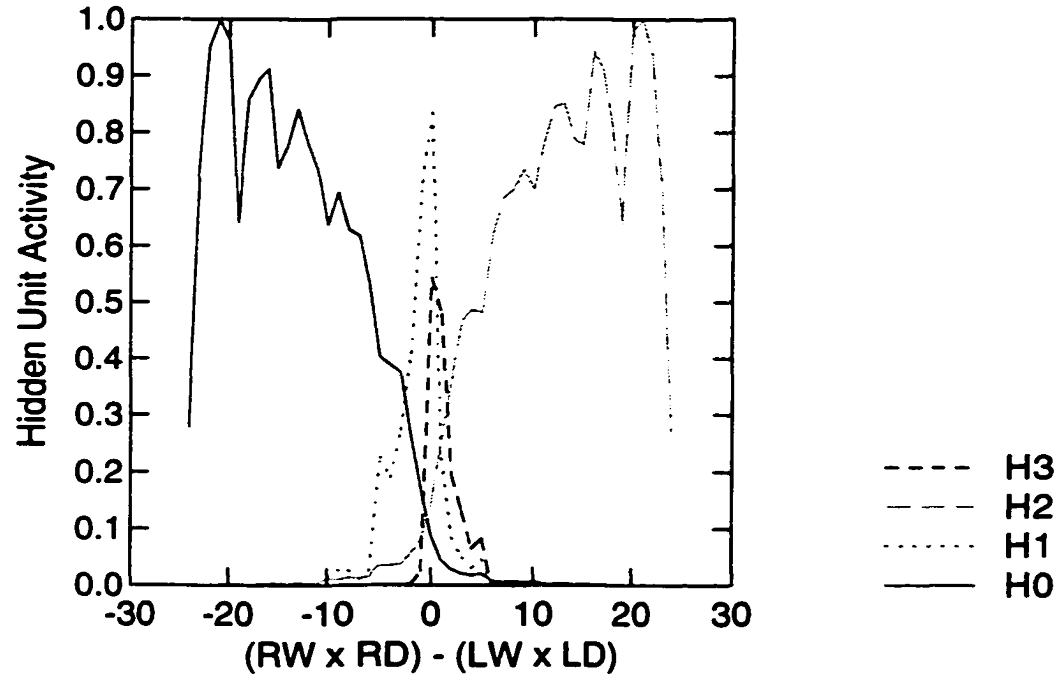

Figure 13. The activity of the four hidden units (H) as a function of the torque algorithm. LD = left distance, LW = left weight, RD = right distance, RW = right weight.

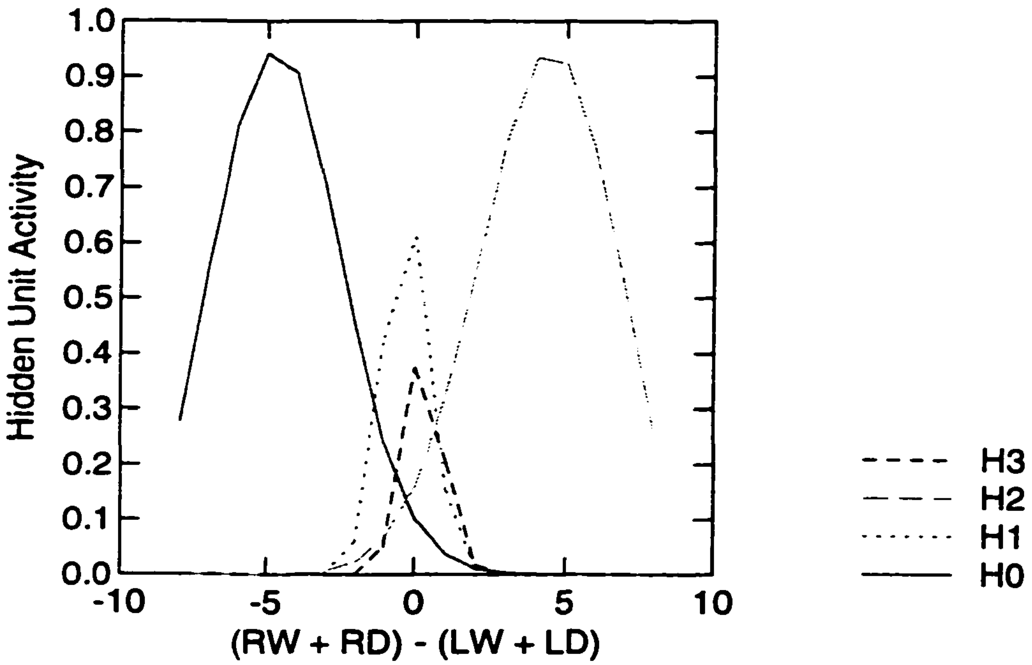

Figure 14. The activity of the four hidden units (H) as a function of the additive equation. LD = left distance, LW = left weight, RD = right distance, RW = right weight.

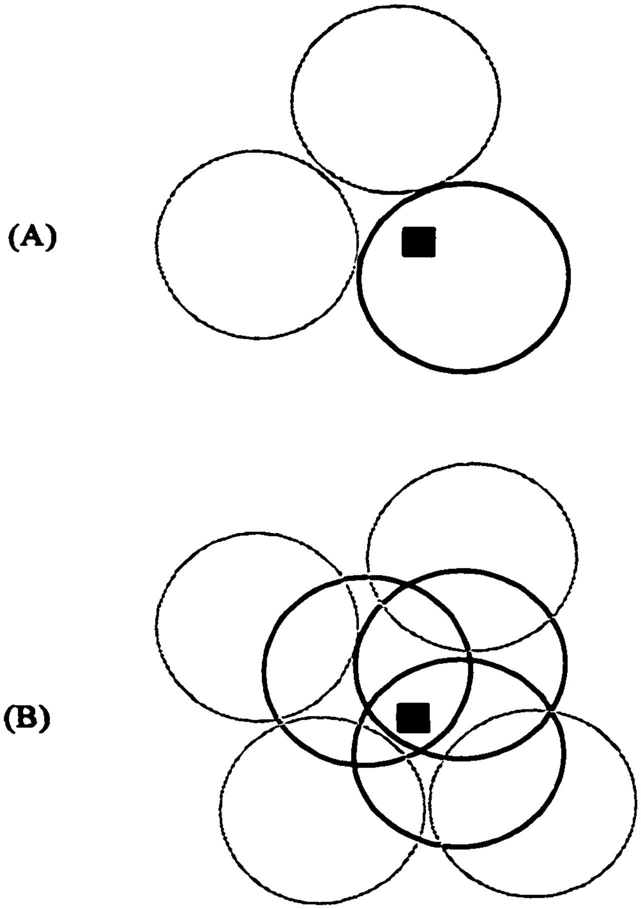

Figure 15. Illustration of the analogy with overlapping receptive fields in the visual system. See text for further explanation.

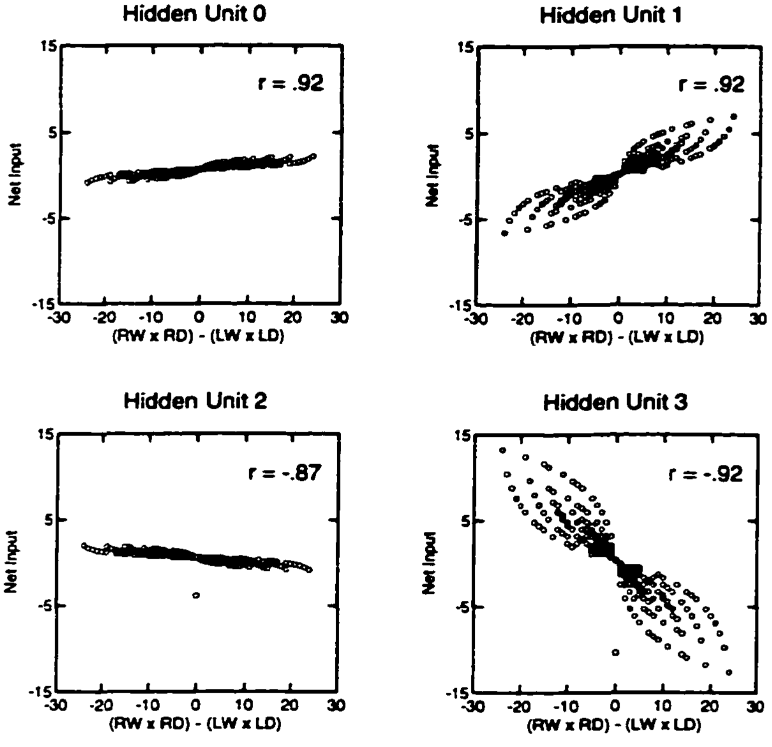

Figure 16. Net input (i.e., the sum of the input units x connection weights from the converged network) as a function of the torque algorithm for the four hidden units in Network 4120. Each dot represents one of the 625 possible balance-scale problems. See text for further explanation.

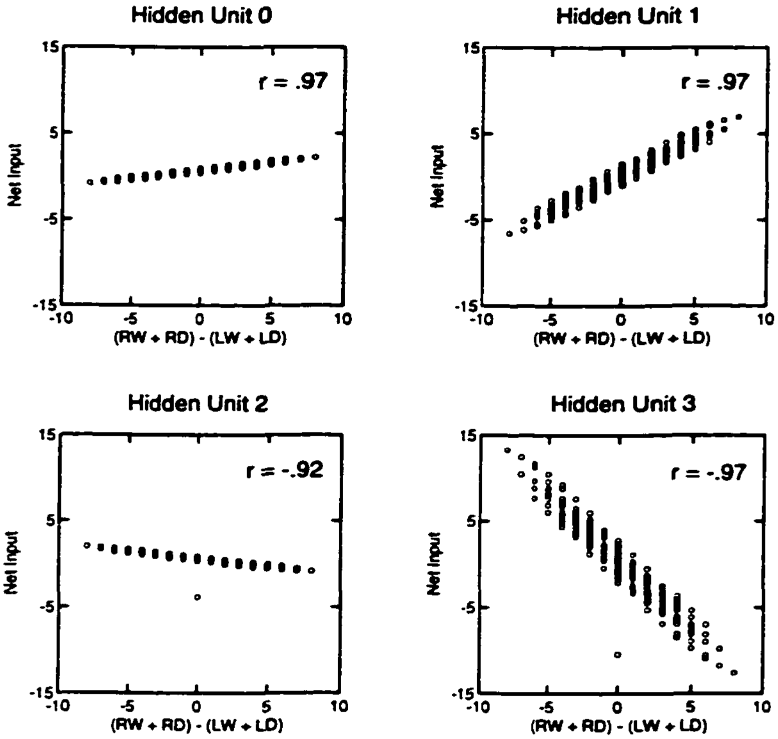

Figure 17. Net input (i.e., the sum of the input units x connection weights from the converged network) as a function of the additive equation [(RW + RD) - (LW + LD)] for the four hidden units in Network 4120.

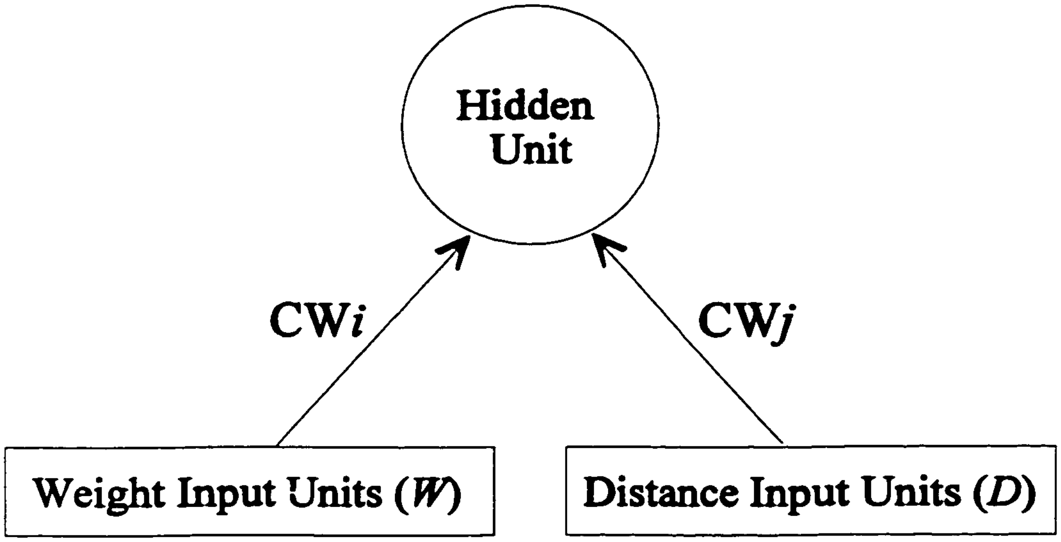

Figure 18. A simplified illustration to demonstrate that multiplication is not a primitive operation of a processing unit (a hidden unit is shown). A single connection is shown between the vector of input units representing weight information (CW*i*) and the vector of information representing distance information (CW*j*). This processing unit cannot multiply the ***weight*** information by the ***distance*** information; it can only sum the information from both sources: *Net Input* = $\sum(\text{CW}i)(W) + \sum(\text{CW}j)(D)$.

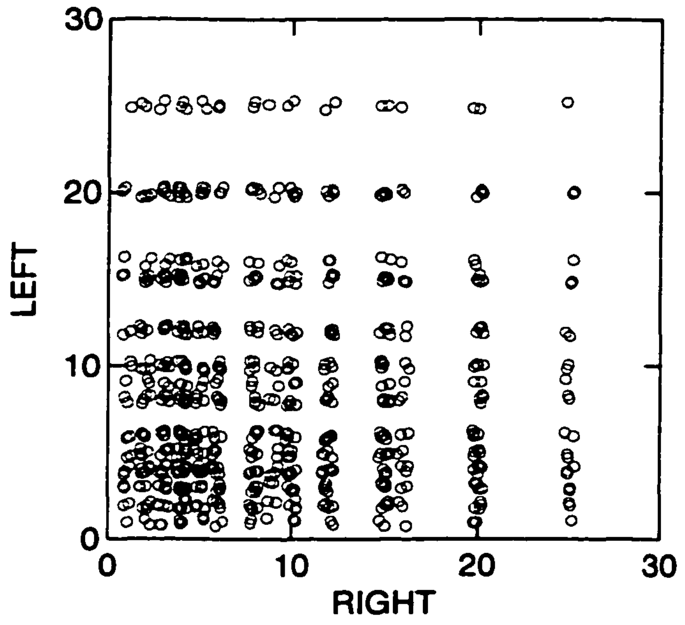

Figure 19. Two-dimensional representation of the entire problem space (625 problems). The four dimensions have been collapsed into two: LEFT (i.e., LW x LD), RIGHT (i.e., RW x RD), resulting in a "torque difference space." The problems are shown with slight random jitter to show overlapping problems.

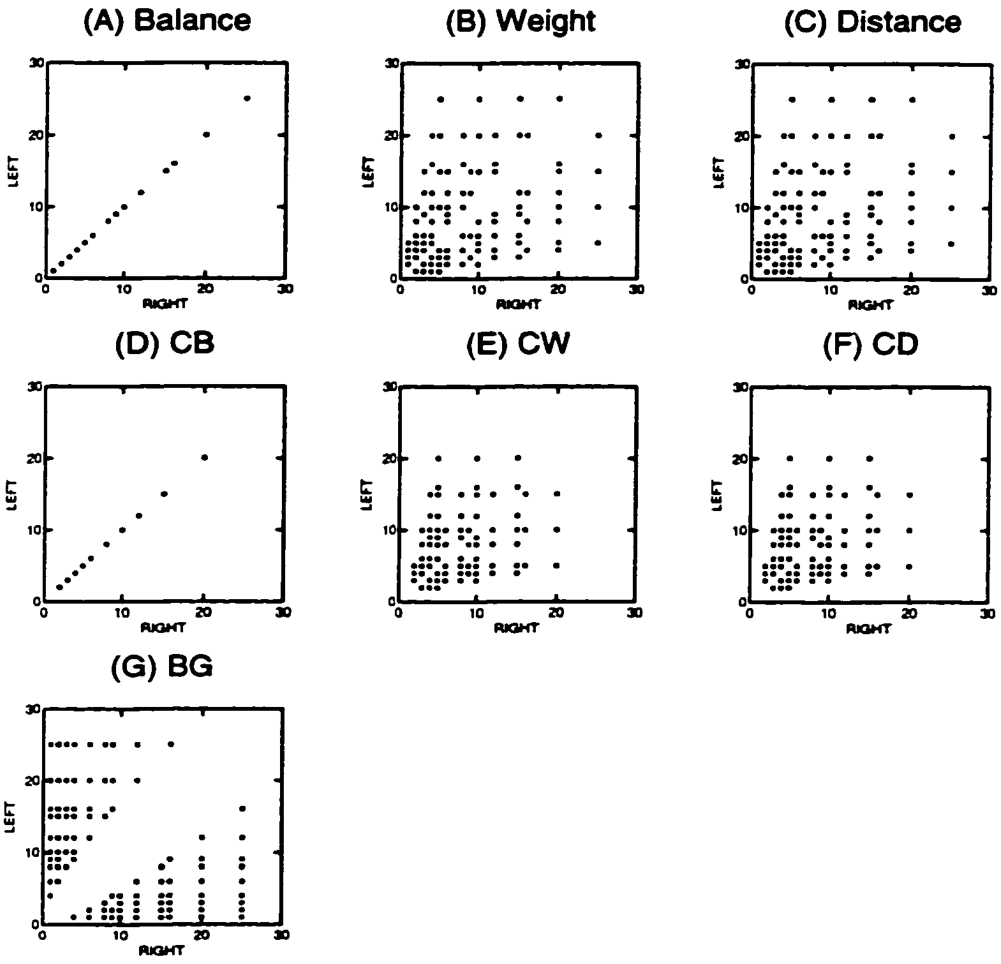

Figure 20. The problem space carved by problem type. Panels A through F are the problem types defined by Siegler (1976). CB = Conflict-Balance, CD = Conflict-Distance, CW = Conflict-Weight, BG = Both-Greater.

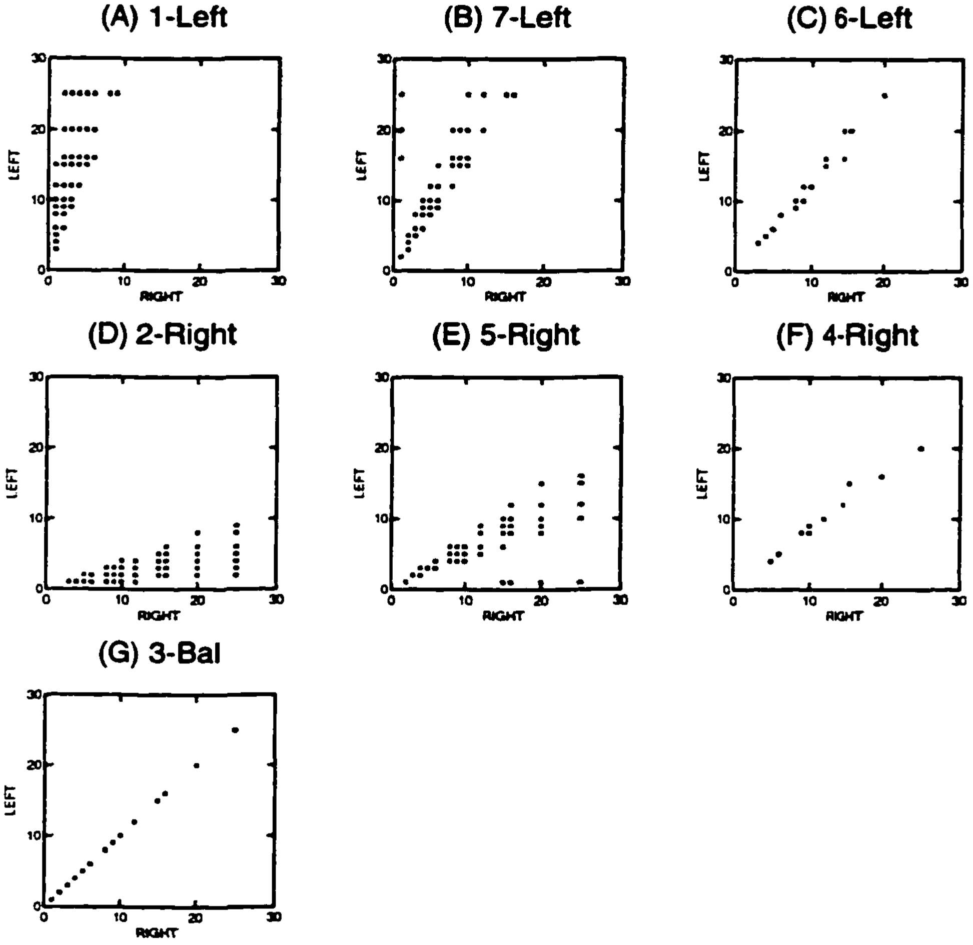

Figure 21. The problem space carved by the cluster analysis of hidden unit activities. Left patterns fall in clusters 1, 6, and 7 (Panels A-C), right patterns fall in clusters 2, 4, and 5 (Panels D-F), and balance problems fall in cluster 3 (Panel G).

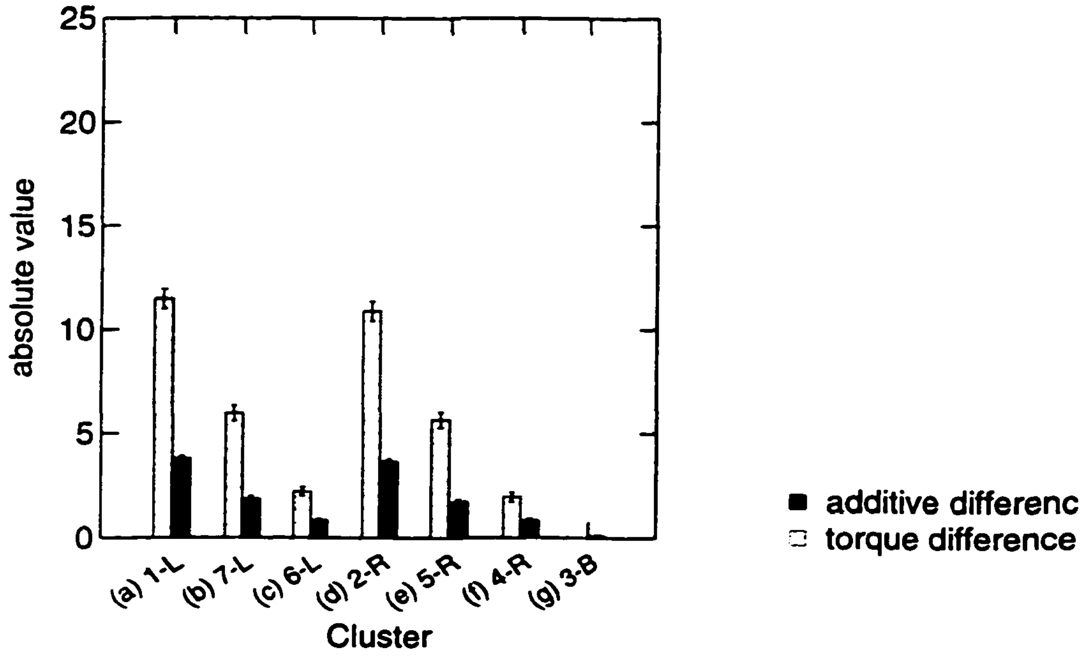

Figure 22. Mean torque difference (absolute value) and mean additive difference (absolute value) for the seven clusters. Torque difference = I RW x RD - LW x LD I. Additive difference = I (RW + RD) - (LW + LD) I. Standard error bars are shown. Numbers refer to the cluster number, B = balance, L = left, R = right.

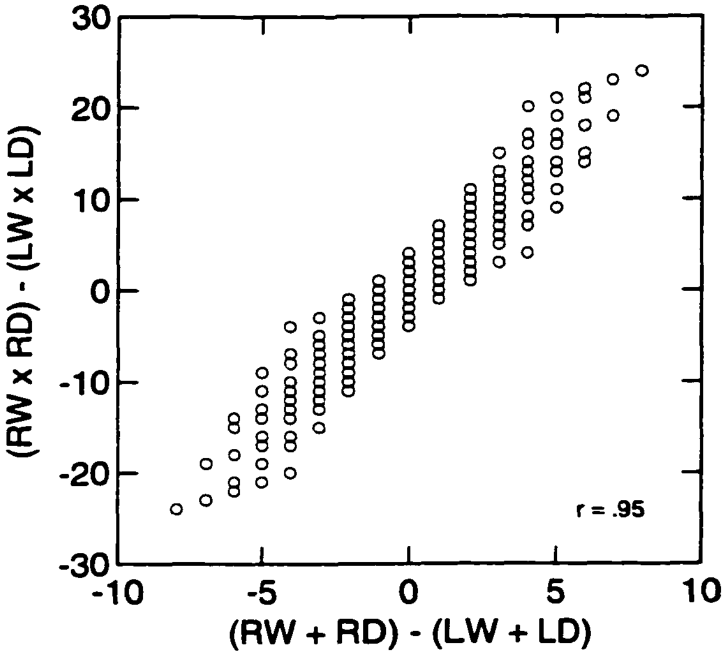

Figure 23. Scatterplot of the relation between the additive equation and the torque equation. LD = left distance, LW = left weight, RD = right distance, RW = right weight.

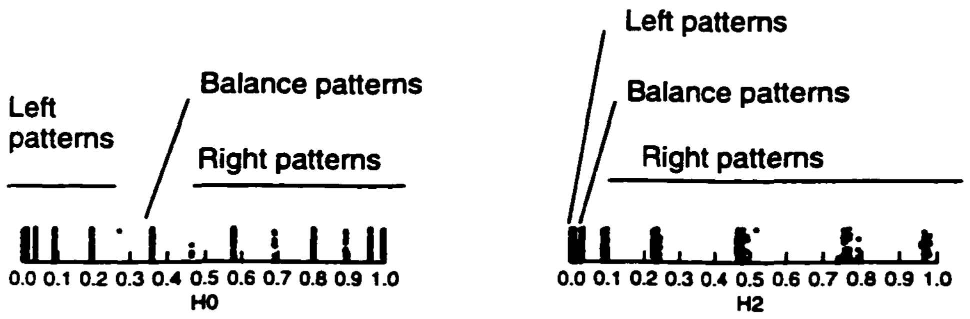

Figure 24. Jittered density plots of hidden unit activity for H0 and H2 for Network 2929 (match problems only). Distinct banding occurs with left, balance, and right patterns falling in distinct regions.

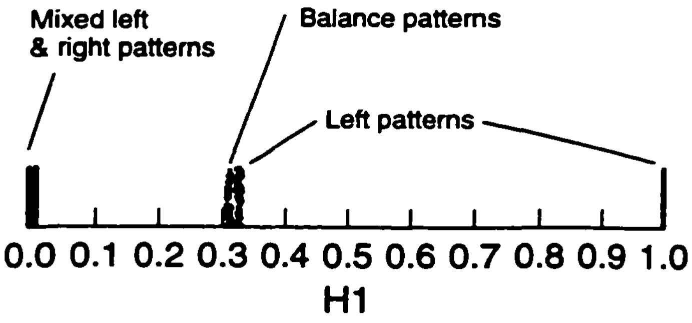

Figure 25. **Jittered density plot of hidden unit activity for H1 for Network 2929.**

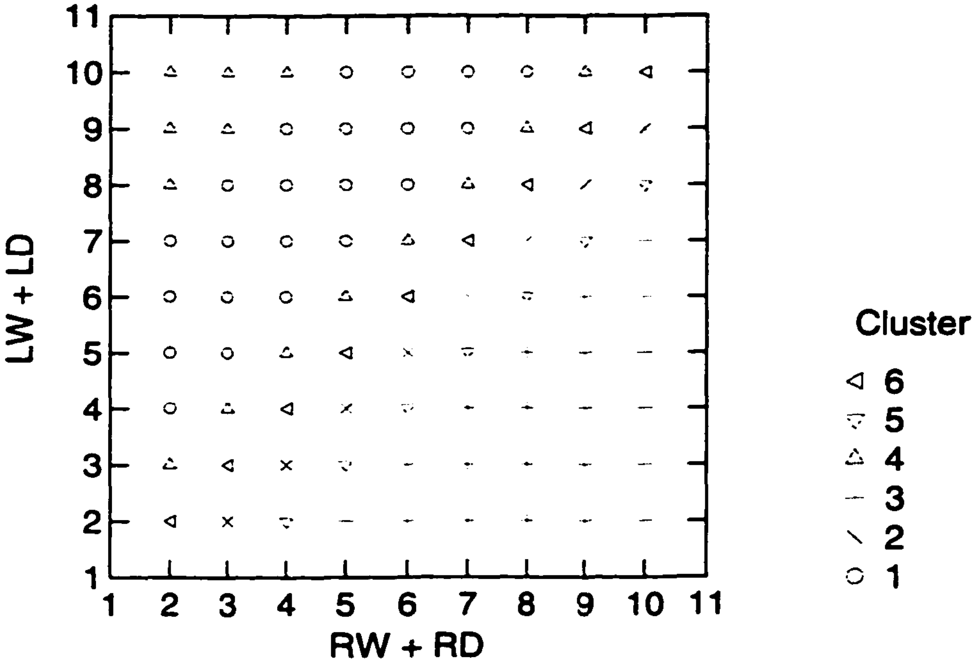

Figure 26. Location of cluster membership for Network 2929 plotted in an "additive difference" space.

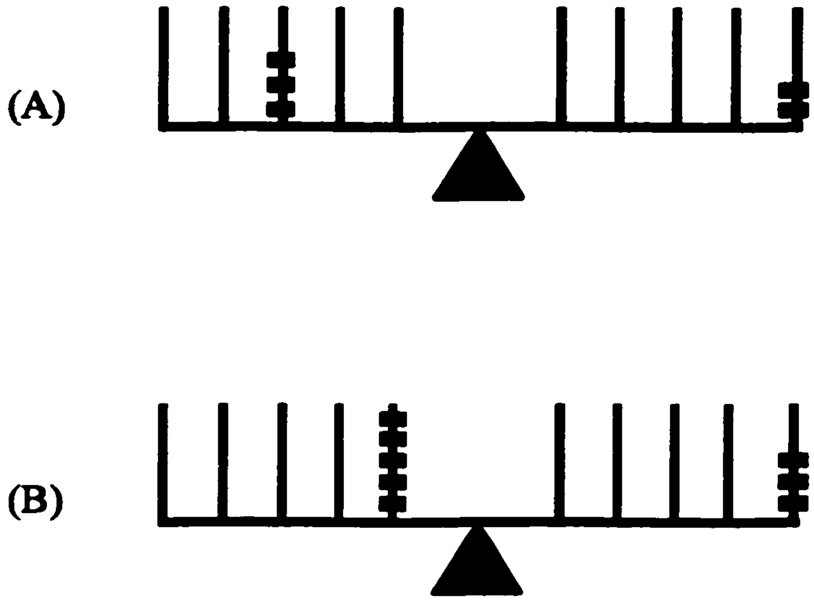

Figure 27. Two conflict-distance problems. The torque difference is 1 in Panel A and 10 in Panel B.

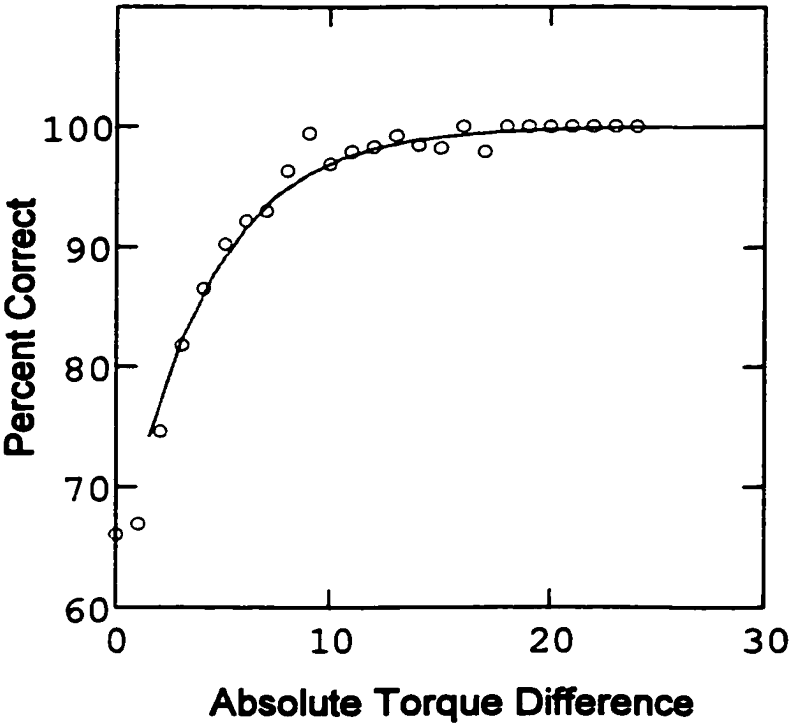

Figure 28. Percent correct plotted as a function of absolute torque difference for all balance-scale problems. The equation of the line is:

percent correct = 100 - 37.660 * EXP(-1 * 0.251 * ABS(TD).

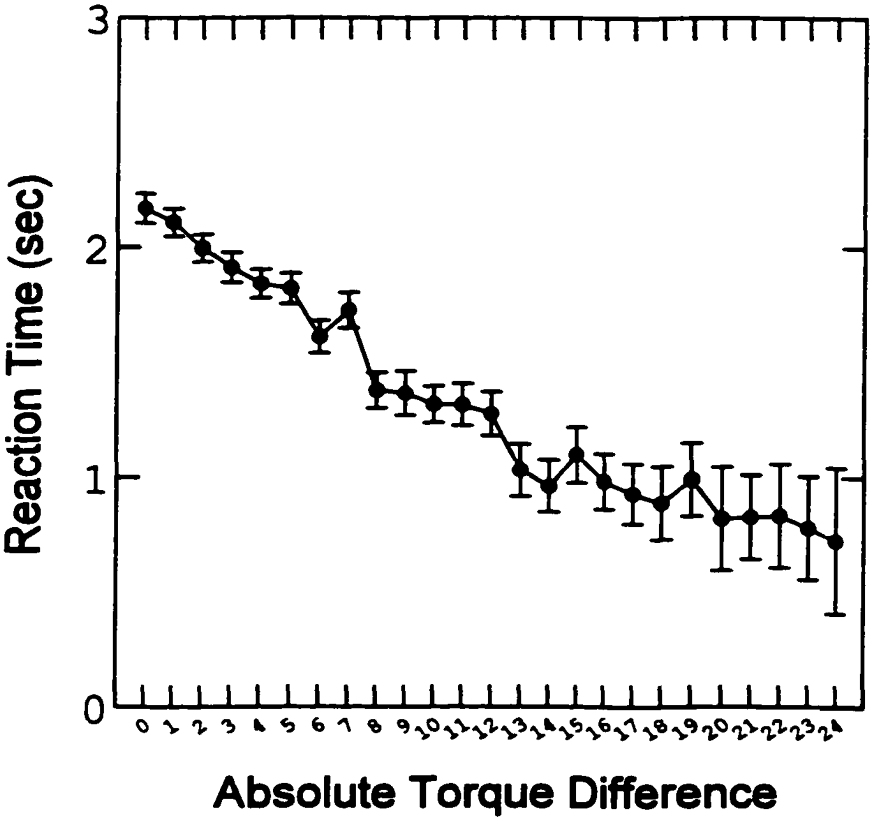

Figure 29. Reaction time as a function of absolute torque difference for all balance-scale problems.

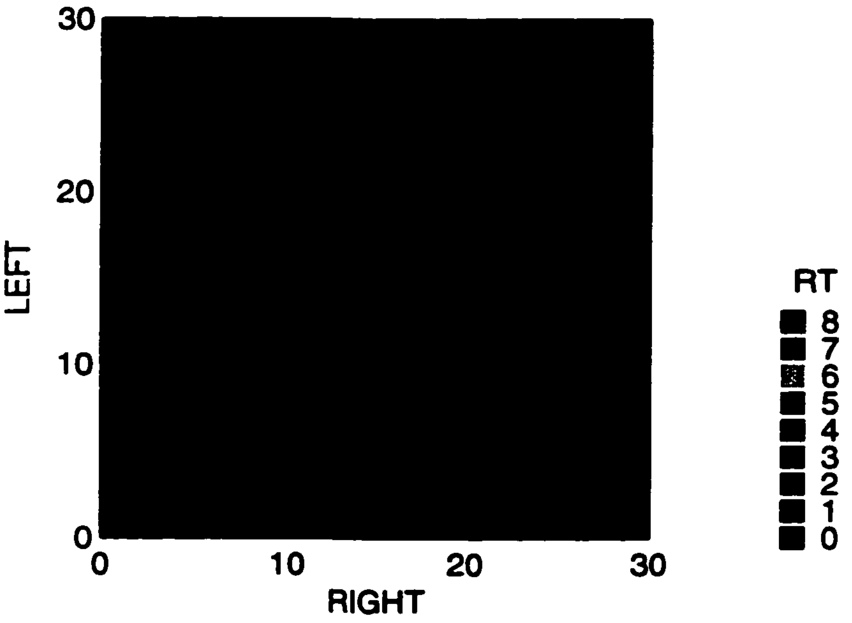

Figure 30. Reaction time for Participant 5 plotted in a 2-D mosaic scatterplot as function of *left torque* (i.e., LW x LD) and *right torque* (i.e., RW x RD). In general, darker shades are associated with faster reaction times.

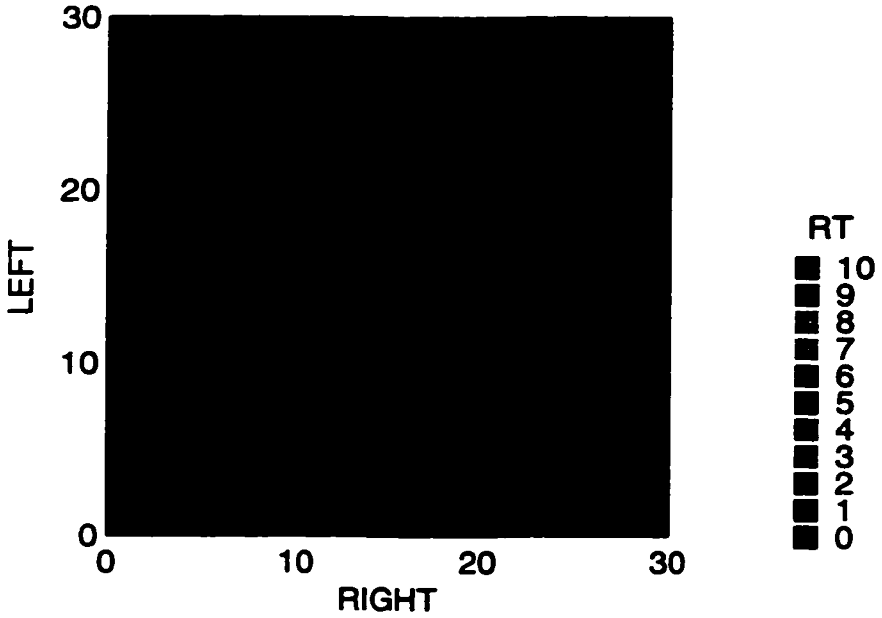

Figure 31. Reaction time for Participant 8 plotted in a 2-D mosaic scatterplot as function of *left torque* (i.e., LW x LD) and *right torque* (i.e., RW x RD). Darker shades are associated with faster reaction times.

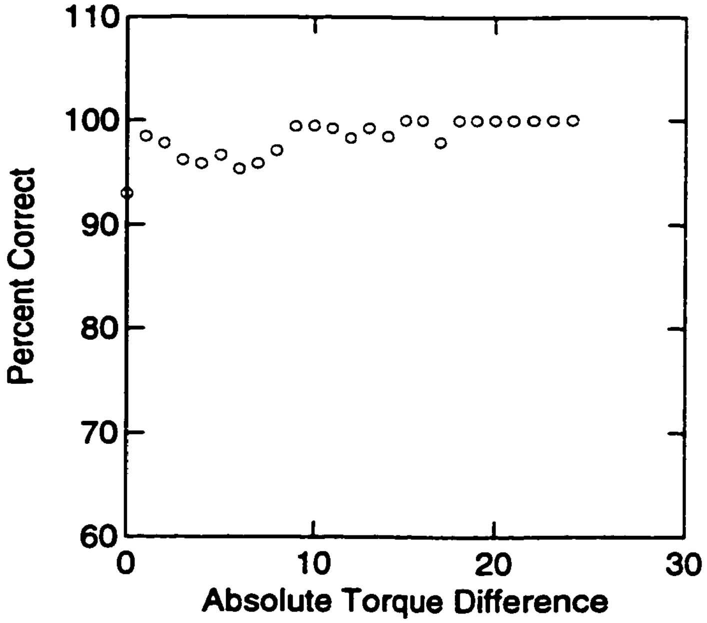

Figure 32. **Percent correct as a function of absolute torque difference for all simple problems (i.e., balance, weight, distance, and both-greater).**

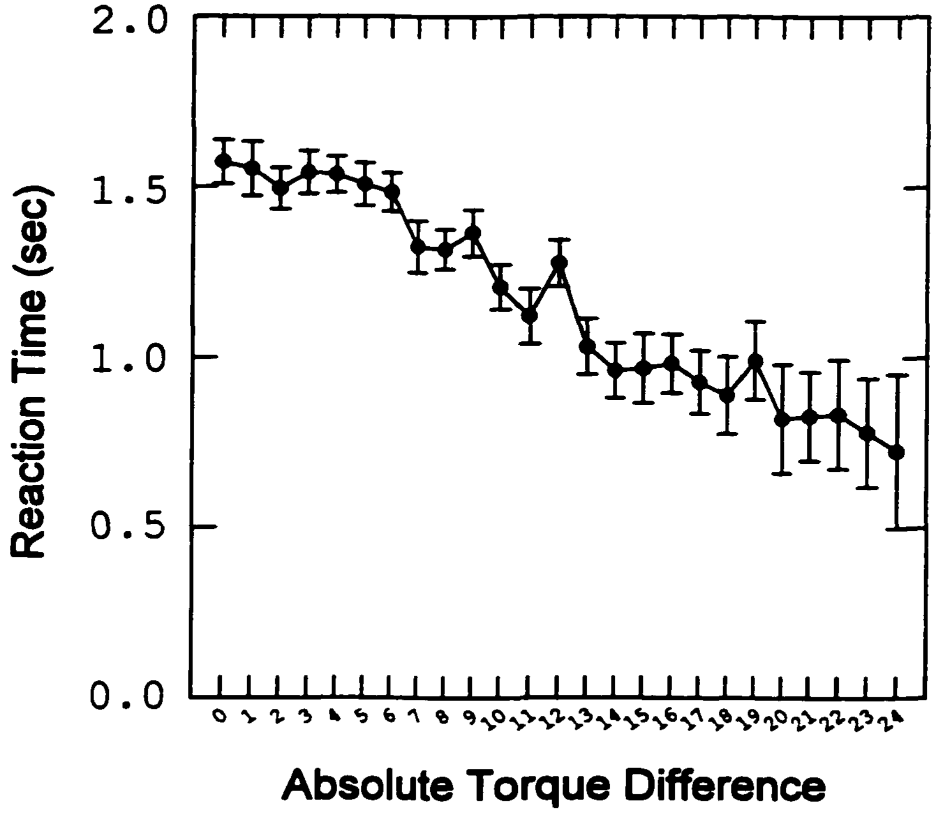

Figure 33. Reaction time as a function of absolute torque difference for all simple problems (i.e., balance, weight, distance, and both-greater).

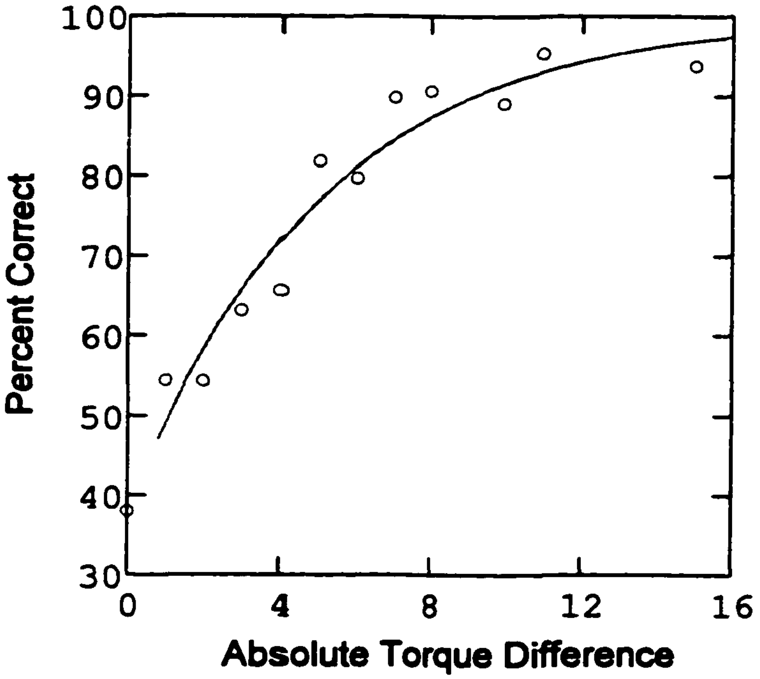

Figure 34. Percent correct as a function of absolute torque difference for all conflict problems. The equation of the line is:

percent correct = 100 - 61.930 * EXP(-1 * 0.199 * ABS(TD)).

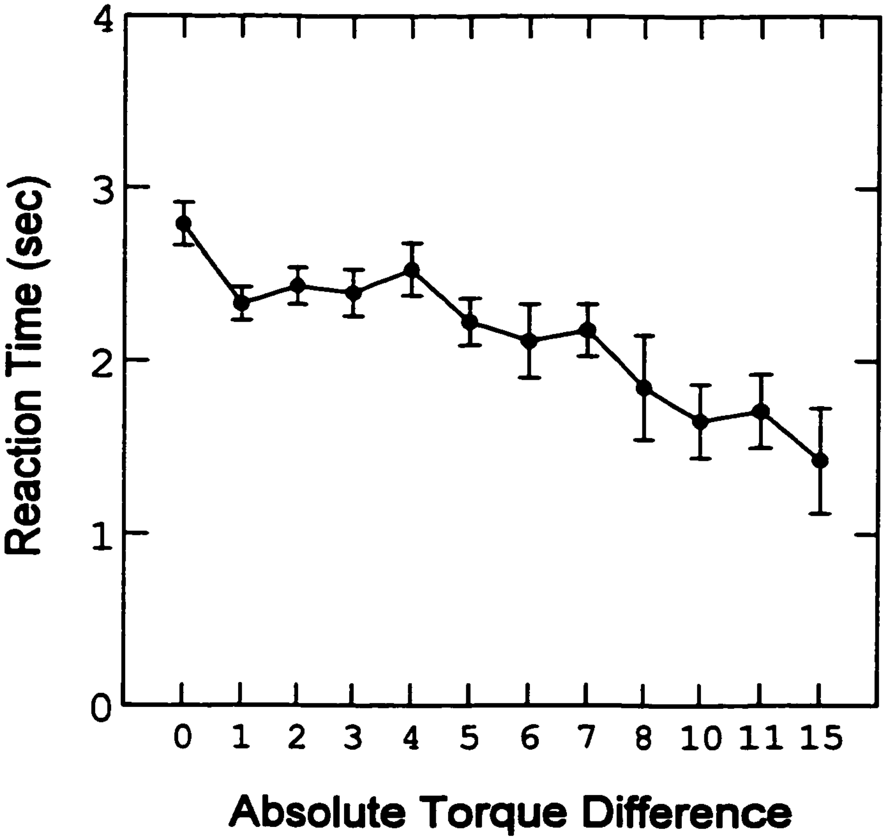

Figure 35. Reaction time as a function of absolute torque difference for all conflict problems.

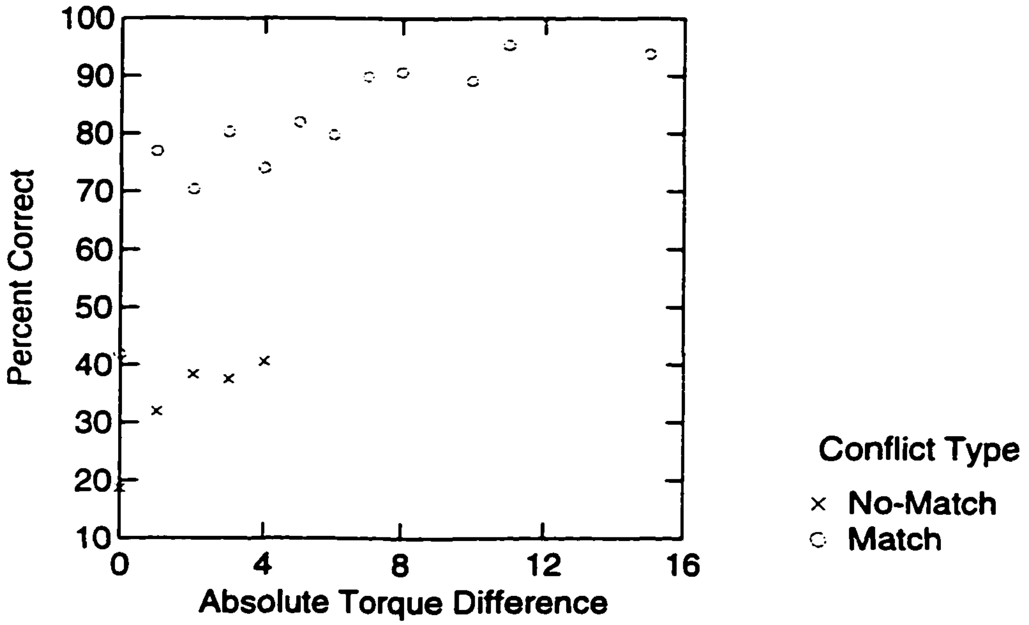

Figure 36. Percent correct as a function of absolute torque difference for no-match and match problems.

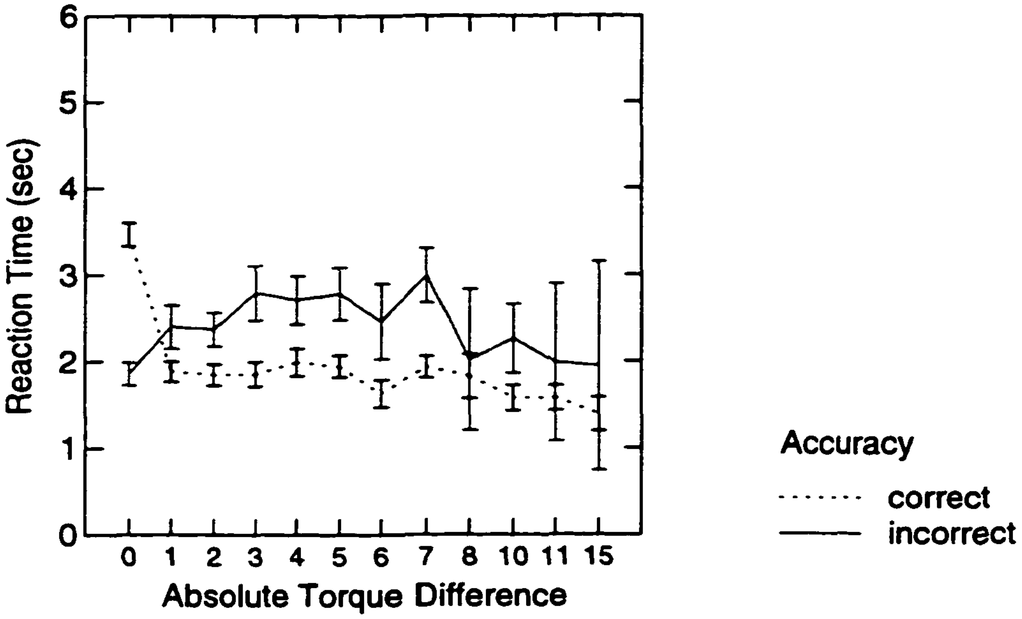

Figure 37. Reaction time as a function of absolute torque difference for *match problems* that were correct and incorrect.

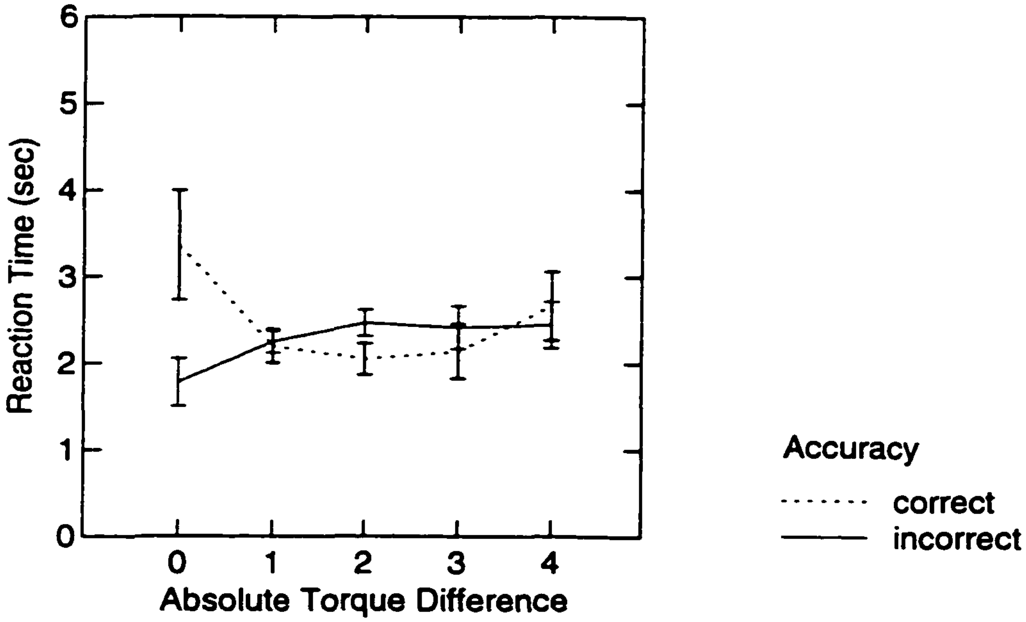

Figure 38. Reaction time as a function of absolute torque difference for *no-match problems* that were correct and incorrect.

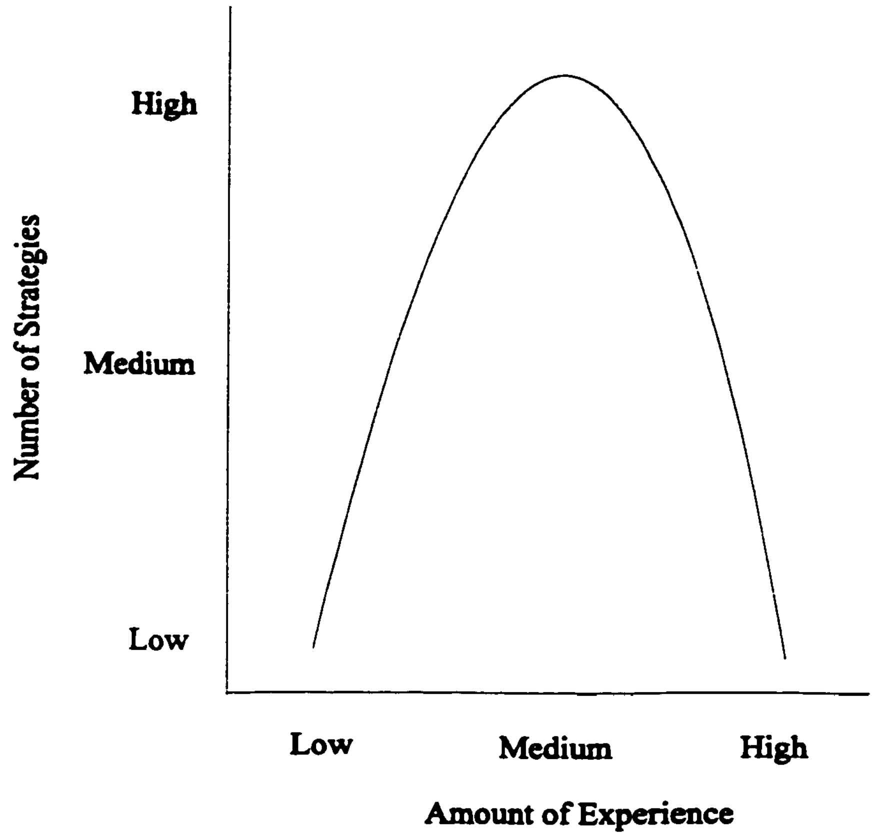

Figure 39. Illustration of Siegler's ***moderate experience hypothesis.*** Variability in strategies is predicted for moderate amounts of experience. Adapted from Siegler (1996).

BIBLIOGRAPHY

Aoki, T. (1991). The relation between two kinds of U-shaped growth curves: Balance-scale and weight-addition tasks. *The Journal of General Psychology*, **118**, 251-261.

Bates, E. A., & Elman, J. L. (1992). *Connectionism and the study of change.* Technical Report #9202, Center for Research in Language, University of California, San Diego, CA.

Bechtel, W., & Abrahamsen, A. (1991). *Connectionism and the mind.* Cambridge, MA: Blackwell.

Bisanz, J., & LeFevre, J. (1990). Strategic and nonstrategic processing in the development of mathematical cognition. In D. F. Bjorkland (Ed.), *Children's strategies: Contemporary views of cognitive development* (pp. 213-244). Hillsdale, NJ: Lawrence Erlbaum.

Berkeley, I. S. N., Dawson, M. R. W., Medler, D. A., Schopflocher, D. P., & Hornsby, L. (1995). Density plots of hidden value unit activations reveal interpretable bands. *Connection Science*, **7**, 167-186.

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.

Dawson, M. R. W. (1998). *Understanding cognitive science.* Oxford, UK: Blackwell.

Dawson, M. R. W., & Medler, D. A. (1996). Of mushrooms and machine learning: Identifying algorithms in a PDP network. *Canadian Artificial Intelligence*, **38**, 14-17.

Dawson, M. R. W., Medler, D. A., & Berkeley, I. S. N. (1997). PDP networks can provide models that are not mere implementations of classical theories. *Philosophical Psychology*, **10**, 25-40.

Dawson, M. R. W., Medler, D. A., McCaughan, D. B., Willson, L., & Carbonaro, M. (1999). Using extra output learning to insert a symbolic theory into a connectionist network: A case study of intertheoretic reduction. Manuscript under editorial review.

Dawson, M. R. W., & Schopflocher, D. P. (1992). Modifying the generalized delta rule to train networks of nonmonotonic processors for pattern classification. *Connection Science*, **4**, 19-31.

Dawson, M. R. W., Willson, L. R., McCaughan, D. B., & Medler, D. A. (1999). A heuristic stopping rule for the k-means cluster analysis of artificial neural networks. Manuscript under editorial review.

diSessa, A. A. (1993). Toward an epistemology of physics. *Cognition and Instruction*, **10**, 105-225.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, **14**, 179-211.

Elman, J. L. (1993). Learning and development in neural networks: The importance of starting small. *Cognition*, **48**, 71-99.

Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking innateness: A connectionist perspective on* *development*. Cambridge, MA: MIT Press.

Fahlman, S. E., & Lebiere, C. (1990). *The cascade-correlation learning architecture*. Technical Report #CMU-CS-90-100, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA.

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. *Journal of Experimental Child Psychology*, **39**, 131-160.

Flavell, J. H. (1985). *Cognitive development* (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Folk, C. L., Egeth, H. E., & Kwak, H. (1988). Subitizing: Direct apprehension or serial processing? *Perception & Psychophysics*, **44**, 313-320.

Hanson, S. J., & Burr, D. J. (1990). What connectionist models learn: Learning and representation in connectionist networks. *Behavioral and Brain Sciences*, **13**, 471-515.

Hinton, G. E., McClelland, J. L., & Rumelhart, D. E. (1986). Distributed representations. In D. Rumelhart, J. McClelland & the PDP Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition., Volume 1: Foundations* (pp. 77-109). Cambridge, MA: MIT Press.

Hinton, G. E., & Sejnowski, T. J. (1986). Learning and relearning in Boltzmann machines. In D. Rumelhart, J. McClelland & the PDP Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition., Volume 1: Foundations* (pp. 282-317).

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.

Johnson, M. (1987). *The body in the mind.* Chicago, IL: University of Chicago Press.

Klahr, D. (1992). Information-processing approaches to cognitive development. In M. H. Bornstein, & M. E. Lamb (Eds.) *Developmental psychology: An advanced textbook (3rd ed.)* (pp. 273-336). Hillsdale, NJ: Lawrence Erlbaum.

Klahr, D., & MacWhinney, B. (1998). Information processing. In D. Kuhn & R. S. Siegler (Eds.), W. Damon (Series Ed.), *Handbook of child psychology (5th ed.): Vol. 2: Cognition, perception, and language.* New York: Wiley.

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* (Vol. 12; pp. 61-116). New York: Academic Press.

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

Kuhn, D., Garcia-Mila, M., Zohar, A., & Andersen, C. (1995). Strategies of knowledge acquisition. *Monographs of the Society for Research in Child*

Development, Serial No. 245, **60**(40), 1-128.

Lavirée, S., Normandeau, S., Roulin, J. L., & Longeot, F. (1987). L'épreuve de la balance de Siegler: Analyse critique du modèle par élaboration de règles [Siegler's balance scale: A critical analysis of the rule-assessment approach]. *L'Année Psychologique*, **87**, 509-534.

Leighton, J. P. (1999). *Reasoning according to the path of least resistance: Constraints from underlying reasoning processes*. Unpublished doctoral dissertation, University of Alberta, Edmonton, AB.

Leighton, J. P., & Dawson, M. R. W. (1999). A PDP approach to understanding Wason's selection task. Manuscript under editorial review.

Mareschal, D., & Shultz, T. R. (1996). Generative connectionist networks and constructivist cognitive development. *Cognitive Development*, **11**, 571-603.

McCaughan, D. B., Medler, D. A., & Dawson, M. R. W. (1999). Internal representations in networks of non-monotonic processing units. *Proceedings of the 1999 International Conference on Neural Networks*. Washington, DC.

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: Oxford University Press.

McClelland, J. L. (1994). The interaction of nature and nurture in development: A parallel distributed processing perspective. In P. Bertelson, P. Eelen, G. d'Ydewalle (Eds.), *International perspectives on psychological science, Volume 1: Leading* *themes*. Hillsdale, NJ: Erlbaum.

McClelland, J. L. (1995). A connectionist perspective on knowledge and development. In T. J. Simon & G. S. Halford (Eds.), *Developing cognitive competence: New approaches to process modeling* (pp. 157-204). Hillsdale, NJ: Lawrence Erlbaum.

McFadden, G. T., Dufresne, A., & Kobasigawa, A. (1987). Young children's knowledge of balance scale problems. *Journal of Genetic Psychology*, **148**, 79-94.

McLeod, P., Plunkett, K., & Rolls, E. T. (1998). *Introduction to connectionist modelling of cognitive processes*. Oxford: Oxford University Press.

Medler, D. A. (1998). *The crossroads of connectionism: Where do we go from here?* Unpublished Doctoral Dissertation, University of Alberta, Edmonton, AB.

Medler, D. A., & Dawson, M. R. W. (1994). Training redundant artificial networks: Imposing biology on technology. *Psychological Research*, **57**, 54-62.

Medler, D. A., McCaughan, D. B., Dawson, M. R. W., & Willson, L. R. (1999). When local isn't enough: Extracting distributed rules from networks. *Proceedings of the 1999 International Conference on Neural Networks*. Washington, DC.

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

Normandeau, S., Lavirée, S., Roulin, J. L., & Longeot, F. (1989). The balance-scale dilemma: Either the subject or the experimenter muddles through. *Journal of Genetic Psychology*, **150**, 237-249.

Ott, L., Larson, R. F., & Mendenhall, W. (1987). *Statistics: A tool for the social* *sciences (4^{th} ed.).* Boston, MA: Duxbury Press.

Perfetti, C. A. (1992). The representation problem in reading acquisition. In P. B. Gough, L. C. Ehri, & R. Treiman (Eds.), *Reading acquisition* (pp. 145-174). Hillsdale, NJ: Erlbaum.

Piaget, J., & Inhelder, B. (1969). *The psychology of the child.* London: Routledge & Kegan Paul.

Plunkett, K., Karmiloff-Smith, A., Bates, E., Elman, J. L., & Johnson, M. H. (1997). Connectionism and developmental psychology. *Journal of Child Psychology & Psychiatry & Allied Disciplines*, **38**, 38-50.

Plunkett, K., & Sinha, C. (1992). Connectionism and developmental theory. *British Journal of Developmental Psychology*, **10**, 209-254.

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

Schauble, L. (1990). Belief revision in children: The role of prior knowledge and strategies for generating evidence. *Journal of Experimental Child Psychology*, **49**, 31-57.

Schauble, L. (1996). The development of scientific reasoning in knowledge-rich contexts. *Developmental Psychology*, **32**, 102-119.

Schauble, L., & Glaser, R. (1990). Scientific thinking in children and adults.

Contributions to Human Development, **21**, 9-27

Schmidt, W. C., & Shultz, T. R. (1991). *A replication and extension of McClelland's balance scale model.* Technical Report 91-10-18. McGill University.

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

Schmidt, W. C., & Ling, C. X. (1996). A decision-tree model of balance scale development. *Machine Learning*, **24**, 203-230.

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

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. *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. J. Simon & G. S. Halford (Eds.), *Developing cognitive competence: New approaches to process modeling* (pp. 205-261). Hillsdale, NJ: Lawrence Erlbaum.

Siegler, R. S. (1976). Three aspects of cognitive development. *Cognitive*

Psychology, **8**, 481-520.

Siegler, R. S. (1978). The origins of scientific reasoning. In R. S. Siegler (Ed.), *Children's thinking: What develops?* (pp. 109-149). Hillsdale, NJ: Lawrence Erlbaum.

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

Siegler, R. S. (1995). Children's thinking: How does change occur? In F. E. Weinert & W. Schneider (Eds.), *Memory performance and competencies: Issues in growth and development* (pp. 405-430). Mahwah, NJ: Lawrence Erlbaum.

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

Siegler, R. S. (1998). *Children's thinking (3rd ed.)*. Upper Saddle River, NJ: Prentice Hall.

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

Siegler, R. S., & Crowley, K. (1991). The microgenetic method: A direct means for studying cognitive development. *American Psychologist*, **46**, 606-620.

Siegler, R. S., & McGilly, K. (1989). Strategy choices in children's time-telling. In I. Levin & D. Zakay (Eds.), *Time and human cognition: A life span perspective*. The Netherlands: Elsevier.

Siegler, R. S., & Shipley, C. (1995). Variation, selection, and cognitive change. In T. J. Simon & G. S. Halford (Eds.), *Developing cognitive competence: New* *approaches to process modeling* (pp. 31-76). Hillsdale, NJ: Lawrence Erlbaum.

Surber, C. F., & Gzesh, S. M. (1984). Reversible operations in the balance scale task. *Journal of Experimental Child Psychology*, **38**, 254-274.

Trick, L. M., & Pylyshyn, Z. W. (1993). What enumeration studies can show us about spatial attention: Evidence for limited capacity preattentive processing. *Journal of Experimental Psychology: Human Perception & Performance*, **19**, 331-351.

Trick, L. M., & Pylyshyn, Z. W. (1994). Why are small and large numbers enumerated differently? A limited-capacity preattentive stage in vision. *Psychological Review*, **101**, 80-102.

van Maanen, L., Been, P., & Sijtsma, K. (1989). The linear logistic test model and heterogeneity of cognitive strategies. In E. E. Roskam (Ed.), *Mathematical psychology in progress* (pp. 267-287). Berlin: Springer-Verlag.

Varnhagen, C. K. (1995). Children's spelling strategies. In V. W. Berninger (Ed.), *The varieties of orthographic knowledge II: Relationships to phonology, reading, and writing* (pp. 251-290). Dordrecht, Netherlands: Kluwer Academic Publishers.

Varnhagen, C. K., McCallum, M., & Burstow, M. (1997). Is children's spelling naturally stage-like? *Reading and Writing: An Interdisciplinary Journal*, **9**, 451-481.

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

Appendix A

The Binomial Formula and Sample Calculation

The binomial formula used for determining the probability of observing *X* successes in *n* number of trials is given by,

$$P(X) = \frac{n!}{X!(n - X)!} p^{X}q^{n-X}$$

where *n* is the number of trials, *X* is the number of successes in *n* trials, *p* is the probability of success in a single trial, and *q* is equal to 1 - *p* (Ott, Larson, & Mendenhall, 1987).

For example, when 4 instances of a problem type are used, the probability of getting 3 correct can be calculated ($n = 4$, $X = 3$). Given that there are three response alternatives, the probability of success on a single trial is 1 in 3 ($p = 1/3$, $q = 2/3$). When we substitute these values into the formula, we get:

$$P(3 \text{ successes}) = \frac{4!}{3!(4 - 3)!} 1/3^{3}2/3^{4-3}$$

$$P(3 \text{ successes}) = \frac{24}{6} (0.037)(0.667)$$

$$P(3 \text{ successes}) = 0.098$$

Therefore, the probability of getting 3 items correct out of 4 by chance alone is 0.098 in the long run.

Appendix B

A Primer of Connectionist Terminology

Several characteristics of connectionist models will be described in order to facilitate the presentation of previous models and the current simulation results. *Structurally*, artificial neural networks (ANN) contain processing units (i.e., input units, hidden units, and output units) and modifiable connection weights between the processing units (see Figure A1). Important *processes* are net input functions, activation functions, output functions, and learning rules. Different types of *architectures* can be constructed from these different structural and processing components.

Processing Units

The neuron is the structural and functional unit of the nervous system (Pinel, 1993). The processing units in a connectionist network are analogous to neurons. Processing units are usually described as one of three types:

Input units. The problem or pattern representation is presented to the network via input units. Typically, the problem or pattern is represented in a distributed manner. For example, balance-scale problems have four pieces of information: left weight, left distance, right weight, and right distance (each with a maximum value of 5). This information could be represented as four inputs (i.e., integer values from 1-5) or as 20 inputs (i.e., 5 input units for each bit of information).

Hidden units. Processing units that do not receive input, or do not represent output, are called hidden units. This is an intermediate layer (or layers) of processing

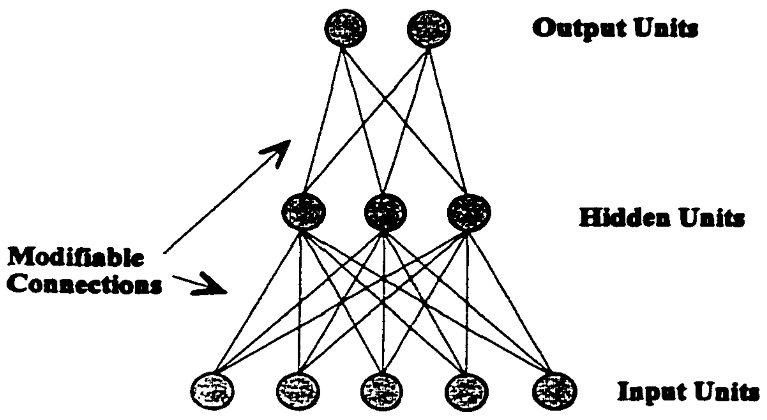

Figure B1. **An example of a simple connectionist network (or multilayer perceptron).**

units that the modeler does not have access to. The role of hidden units is to detect features in the input patterns that allow the network to make a correct response on the output units.

Output units. The output unit represents the answer or solution to the problem. Typically, outputs are also distributed representations. In the case of the balance-scale task, for example, there are three possible solutions (tip left, tip right, or balance).

Modifiable connections. Between each layer of processing units are modifiable, weighted connections. These connections function as the means of communication between processing units. The nature (e.g., excitatory or inhibitory) and strength of the connection is determined by its weight. Note that for some types of network architecture (i.e., the multilayer perceptron), connections do not exist between processing units within a particular layer (see Figure B1).

Processing Functions

Net input functions. Hidden units and output units must process the collective activations received from all connections. The signal that enters any given unit *i* from another unit *j* is the signal sent from unit *j* multiplied by the value of the connection weight, w_{ij}. Given that there are several connections to any unit *i*, the processing unit must have some way of computing the total signal it receives. A very common input function is one that simply sums the weighted signals from all the units connected to it, or:

$$\text{Net input}_i = \sum(\text{output}_j)(w_{ij})$$

Activation functions. The activation function calculates the internal level of activity in the processing unit. There are many different activation functions (see Figure B2). The *step function* is illustrated in Panel A. Here, if the net input is less than a particular threshold, the activation of the unit is 0; if the net input is greater than this threshold, the activation is 1. The *logistic function* is illustrated in Panel B (also called a sigmoidal activation function). A negative input results in an activation between 0 and 0.5, an net input of 0 results in an activation of 0.5, and positive values result in an activation of greater than 0.5. Net input values of 0 result in the maximum activation value of 1 when the *Gaussian activation function* is used (see Panel C).

Output functions. Once a unit has processed the information from all connecting units, it must propagate a signal to the next layer in the case of hidden units (i.e., to another layer of hidden units, or to the output units), or produce an output response in the case of output units. In most cases, the output function is simply the identity function--the signal sent out by a unit is the unit's level of activity as determined by the activation function.

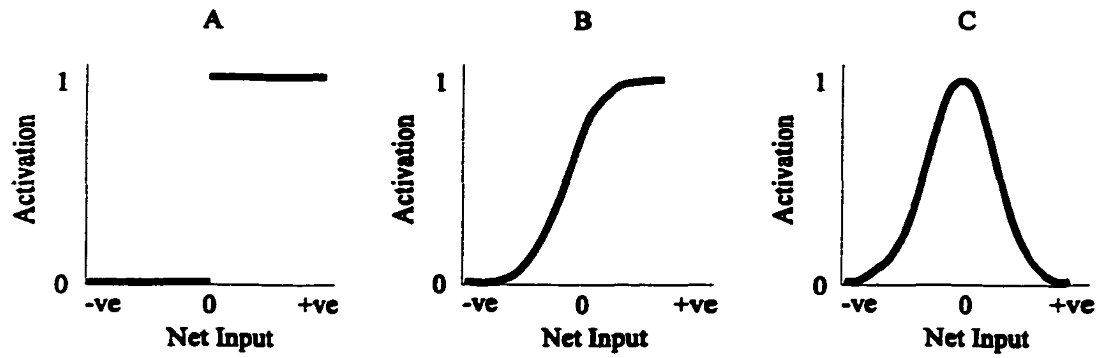

Figure B2. Three common activation functions: (A) the step function, (B) the logistic function, and (C) the Gaussian function. Adapted from Dawson (1998).

Learning algorithms. There are many different learning rules that can be used to train an artificial neural network. Common learning algorithms are (a) the Hebb rule; (b) the delta rule, (c) the generalized delta rule and (d) the Quickprop rule.

A general description of how learning occurs is as follows. Input patterns are presented to the network, and the activity is propogated in a forward manner to the hidden and output units. The network then compares the ***actual*** response to the ***desired*** response (i.e., the ***error***). This error term is propogated backwards through the network, such that the connection weights are changed in order to reduce the error. Training occurs in this way until some criterion of error reduction has occurred.

Appendix C

Characteristics of the *No-Match* Problems

LW	LD	RW	RD	Tip	Problem Type	Torque Difference [a]	Additive Difference [b]	Cluster
1	4	2	2	B	CBSD	0	-1	3
2	2	1	4	B	CBSD	0	1	3
4	1	2	2	B	CBSW	0	-1	3
2	2	4	1	B	CBSW	0	1	3
3	3	5	1	L	CDSB	-4	0	7
2	4	5	1	L	CDSB	-3	0	7
4	2	5	1	L	CDSB	-3	0	7
2	3	4	1	L	CDSB	-2	0	7
3	2	4	1	L	CDSB	-2	0	7
3	4	5	2	L	CDSB	-2	0	6
4	3	5	2	L	CDSB	-2	0	6
2	2	3	1	L	CDSB	-1	0	6
3	3	4	2	L	CDSB	-1	0	6
4	4	5	3	L	CDSB	-1	0	6
2	3	5	1	L	CDSW	-1	1	6
3	2	5	1	L	CDSW	-1	1	6
3	3	1	5	L	CWSB	-4	0	7
2	4	1	5	L	CWSB	-3	0	7
4	2	1	5	L	CWSB	-3	0	7
2	3	1	4	L	CWSB	-2	0	7
3	2	1	4	L	CWSB	-2	0	7
3	4	2	5	L	CWSB	-2	0	6
4	3	2	5	L	CWSB	-2	0	6
2	2	1	3	L	CWSB	-1	0	6
3	3	2	4	L	CWSB	-1	0	6

4	4	3	5	L	CWSB	-1	0	6
2	3	1	5	L	CWSD	-1	1	6
3	2	1	5	L	CWSD	-1	1	6
3	1	2	2	R	CDSB	1	0	5
4	2	3	3	R	CDSB	1	0	4
5	3	4	4	R	CDSB	1	0	4
4	1	2	3	R	CDSB	2	0	5
4	1	3	2	R	CDSB	2	0	5
5	2	3	4	R	CDSB	2	0	4
5	2	4	3	R	CDSB	2	0	4
5	1	2	4	R	CDSB	3	0	5
5	1	4	2	R	CDSB	3	0	5
5	1	3	3	R	CDSB	4	0	5
5	1	2	3	R	CDSW	1	-1	4
5	1	3	2	R	CDSW	1	-1	4
1	3	2	2	R	CWSB	1	0	5
2	4	3	3	R	CWSB	1	0	4
3	5	4	4	R	CWSB	1	0	4
1	4	2	3	R	CWSB	2	0	5
1	4	3	2	R	CWSB	2	0	5
2	5	3	4	R	CWSB	2	0	4
2	5	4	3	R	CWSB	2	0	4
1	5	2	4	R	CWSB	3	0	5
1	5	4	2	R	CWSB	3	0	5
1	5	3	3	R	CWSB	4	0	5
1	5	2	3	R	CWSD	1	-1	4
1	5	3	2	R	CWSD	1	-1	4

Note. B = balance, L = left tip, R = right tip, LD = left distance, LW = left weight, RD = right distance, RW = right weight, CB = conflict balance, CD = conflict distance, CW = conflict weight

[a] Torque difference = LW x LD - RW x RD. [b] Additive difference = (LW + LD) - (RW + RD).

Appendix D

Number of Problems at Each Level of Torque Difference for a Six-Peg, Six-Weight Problem Set for Non-Balance Problem Types [a]

Torque	Problem Type				
Difference [b]	CD	CW	DIS	WT	Torque Difference Level
1	24	24	10	10	1 (simple & conflict)
2	32	32	18	18	
3	22	22	16	16	2 (simple & conflict)
4	20	20	22	22	
5	10	10	12	12	3 (conflict)
6	20	20	24	24	
7	12	12	0	0	
8	10	10	12	12	
9	8	8	6	6	
10	8	8	10	10	
11	4	4	0	0	
12	6	6	18	18	3 (simple)
13	4	4	0	0	
14	8	8	0	0	
15	2	2	8	8	
16	0	0	4	4	
18	4	4	6	6	4 (conflict)
19	4	4	0	0	4 (conflict)
20	0	0	6	6	
24	2	2	4	4	4 (simple & conflict)
25	0	0	2	2	4 (simple)
30	0	0	2	2	4 (simple)

Note. Problems underlined were sampled for test sets used by Ferretti and Butterfield (1986). CD = conflict-distance, CW = conflict-weight, DIS = distance, WT = weight. Some values of TD are not shown as there are no problems in those categories (i.e., 17, 21-23, 26-29). For simple problems (WT and DIS), the four levels of torque difference (TD) were 1, 3, 12, and 24-30. For conflict problems, the four levels of TD were 1, 3, 5, and 18-24. It is unclear why at levels 3 and 4 there is such a large disparity in TD for simple and conflict problems, given that problems of comparable TD exist.

[a] Excluding both-greater problems. [b] Absolute value of (RW x RD - LW x LD).

Appendix E

Number of Problems at Each Level of Torque Difference for a Five-Peg, Five-Weight Problem Set for Non-Balance Problem Types [a]

Torque	Problem Type				
Difference [b]	CD	CW	DIS	WT	Torque Difference Level
1	20	20	8	8	1 (simple & conflict)
2	16	16	14	14	
3	10	10	12	12	2 (simple & conflict)
4	8	8	16	16	
5	10	10	8	8	3 (conflict)
6	4	4	10	10	
7	8	8	0	0	
8	2	2	8	8	
9	0	0	4	4	
10	4	4	6	6	4 (conflict)
11	4	4	0	0	4 (conflict)
12	0	0	6	6	3 (simple)
13	0	0	0	0	
14	0	0	0	0	
15	2	2	4	4	4 (simple & conflict)
16	0	0	2	2	4 (simple)
17	0	0	0	0	
18	0	0	0	0	
19	0	0	0	0	
20	0	0	2	2	4 (simple)

Note. Problems underlined were sampled for test sets in Schmidt and Shultz (1991). For simple problems (WT and DIS), the four levels of torque difference (TD) were 1, 3, 12, and 15-20. For conflict problems, the four levels of TD were 1, 3, 5, and 10-15. Notice that the TD at level 3 for simple problems is larger than the TD for some conflict problems at level 4. CD = conflict-distance, CW = conflict-weight, DIS = distance, WT = weight.

[a] Excluding both-greater problems. [b] Absolute value of (RW x RD - LW x LD).

Appendix F

Cross-tabulation of Cluster Membership (Network 4120) and Problem Type

Cluster	BAL	WT	DIS	BG	CW Match	CW NM	CD Match	CD NM	CB Match	CB NM
6 (50)	–	10	10	–	8	5[a] 2[b]	8	5[c] 2[d]	–	–
7 (121)	–	25	25	25	18	5[a]	18	5[c]	–	–
1 (117)	–	15	15	75	6	–	6	–	–	–
3 (49)	25	–	–	–	–	–	–	–	20	4[e]
4 (32)	–	5	5	–	5	4[a] 2[b]	5	4[c] 2[d]	–	–
5 (129)	–	28	27	24	19	6[a]	19	6[c]	–	–
2 (127)	–	17	18	76	8	–	8	–	–	–

Note. Clusters 1, 6, 7 contain left problems, clusters 2, 4, 5 contain right problems. Number of instances are shown in brackets. BAL = balance, WT = weight, DIS = distance, CB = conflict-balance, CD = conflict-distance, CW = conflict-weight, NM = no-match.

[a] Conflict weight sum balance (CWSB). [b] Conflict weight sum distance (CWSD).

[c] Conflict distance sum balance (CDSB). [d] Conflict distance sum weight (CDSW).

[e] Conflict balance sum weight (CBSW), Conflict balance sum distance (CBSD).

Appendix G

C++ Code to Measure Reaction Time

```
void query(void)  // Collecting Responses and Reaction Times
{
  clock_t start, end;            // initialize a timer or clock here
  printf("\nGiven this configuration of weights and pegs, will this balance scale");
  printf("\n\t(a) Tip to the LEFT? ...if so press the 1 key");
  printf("\n\t(b) Balance? ...if so press the 2 key");
  printf("\n\t(c) Tip to the RIGHT? ...if so press the 3 key\n");
  printf("\nPlease make your selection >");
  start = clock();               // start a timer or clock here
  for (;;)                       // will only accept one of 3 keystrokes
  {                              // waits until one is pressed
    temp=getche();
    if (temp=='1' || temp=='2' || temp=='3')
    {break;}
  }
  end = clock();
  float rt;
  rt = ((end - start) / CLK_TCK ) ;   //macro in <time.h>; expresses in sec. to 3 dec. places
  fprintf(output,"%d %c %F", BLOCK, temp, rt);  //adding to the output file

  // Initiate a new trial (end query)
  sleep(1);
  clrscr();
  printf("\n\n\n\n\tTo start the next trial, hit the SPACEBAR");
  wait = 0;
  while(wait != ' ')
  {
    while (kbhit())
        {
        wait = getche();
        }
  }
}  //end of query function
```

Appendix H

Written Protocols from Participants in Experiment 3

After completing the task, participants were asked to write responses to three questions: (a) Are you familiar with the concept "torque"?, (b) Can you define it?, and (c) When doing the task, did you have any particular strategy that you can describe?

Participant 1: No; No; "I remembered being on a teeter-totter with my younger brother and I had to sit closer to the middle so I tried to adjust for how far out the weights were."

Participant 2: No; No; "Blocks toward the outer pegs were "heavier" than that same amount of weight closer to the middle. More weight on one side meant that balance should fall toward that side, unless more weight was closer to the middle, and weight on the other side (less weight) was toward the outside (balance, or fall towards side with weights closer to the outside)."

Participant 3: Yes; No; "Balanced distance from center compared to mass" (diagram of simple distance problem; [under side with greater distance] "pulls down more as has [more] torque."

Participant 4: No; No; "Nothing special, just common sense."

Participant 5: No; No; "I looked at the side that had more blocks and then what peg it was on and then at the other side to see what peg the other blocks were on and whatever side had the peg closer to the end I thought was the side it tipped to."

Participant 6: No; No; "Visualizing how the balance would react, using a counterbalance further in than out. (e.g., lw=5, ld=1, rw = 1, rd=5) as balancing, then trying to consistently use the same principle."

Participant 7: No; No; "Trying to equalize distance and weight, although I had no formal rule that I could use quickly enough."

Participant 8: Yes; "to fine tune something"; "I tried to imagine what would be equivalent e.g., if one weight on outer stake equivalent to 2 weights on the next stake in on the other side."