INTRODUCTION

The editors of this book asked the contributors to answer three basic questions about cognitive development: what, how and why. We will attempt to answer these questions by a careful study of one famous case of cognitive development: Piaget's balance scale task for assessing proportional reasoning.

In the past, we have investigated this task in various ways and we will take the present opportunity to summarize and integrate our main findings. We approach the study of cognitive development from a methodological point of view. In each of the forthcoming sections, we apply a new technique in order to get a new perspective on the developmental process.

We first shortly discuss the background of research on the balance scale task. In the second section we attempt to resolve the criticism of Siegler’s (1981) rule theory, the main theory for balance scale reasoning, by introducing categorical latent structure models for rule assessment. In the third section we attempt to validate and extend this rule theory with response time (RT) measures. Using RTs allowed the test of the rule theory in great detail and lead to a number of specifications and improvements of the theory. Fourth, we summarize our study of the transitions between the rules. Knowledge of these transitions is an important step in understanding the mechanisms involved in the development of
proportional reasoning. We focus on a particular transition, from Rule I to Rule II, to investigate whether this transition is a genuine phase transition. Based on our results, we present a restricted overlapping waves model of proportional reasoning, in which some transitions are sudden and others are more continuous. We will describe an ACT-R model that implements the model and thereby partly explains the what, how and why of cognitive development.

HISTORY OF RESEARCH ON THE BALANCE SCALE TASK

In the balance scale task, children are asked to predict the movement of a balance scale. Equally heavy weights can be placed on the arms of the scale, at equally spaced distances. Figure 1 shows a graphical display of a balance scale problem. Blocks underneath the arms prevent the scale from tipping. The task assesses proportional reasoning as it requires the understanding of the multiplicative relation between the dimensions.

Figure 1. Example of a balance scale item.

Inhelder & Piaget (1958) formulated a developmental course of performance on the task, which is characterized by several qualitatively distinct, increasingly complex, stages. The stages range from considering the number of weights only to comparing the products of weight and distance of both sides of the scale (torque-rule). However, stages can not be observed directly. Although responses to balance scale problems are probably related to the developmental stages, the relation is not perfect as both false positive errors (e.g., when
children incidentally derive the correct response by guessing) and false negative errors (e.g.,
when children err) may occur.

Mainly in the context of another Piagetian task, conservation, researchers have asked
what criteria need to be fulfilled to determine that a child possesses true knowledge (Braine
& Shanks, 1965; Brainerd, 1973; Gruen, 1966; Inhelder, Bovet, Sinclair & Smock, 1966;
Smedslund, 1963, 1969; Smith, 1992, 1993). Researchers have argued that a clearer picture
of children’s true knowledge is derived from children’s explanations of their responses.
Supporters of this “Judgment & justification-view” believe that the probability of a false
positive error decreases when an explanation is asked. Supporters of the “Judgment-only
view” reasoned that both types of errors occur with a verbal method, as there is no one-to-one
correspondence between logic and action nor between logic and language. A false positive
error occurs if a child, who does not understand the principle of torque, is taught the correct
justification for the balance scale task. However, when the child is presented with a slightly
different problem, the child will not be able to give the correct justification. False negative
errors may occur because children are not always able to verbally express their mental
operations. Boden (1994) notes that knowledge must exist in action before it can be
verbalized. Hence, a verbal justification may be a sufficient, but not a necessary, condition.
Brainerd (1973) claims that asking for verbal justifications even increases the probability of
false negative errors. According to Brainerd, asking for judgments alone does not necessarily
cause an increased probability of false positive errors.

Research on the complex social interaction between experimenter and child during the
verbal task (Bijstra, van Geert & Jackson, 1989; Donaldson, 1978; Elbers, 1989; McGarrigle
& Donaldson, 1975; Rose and Blank, 1974) and the linguistic aspects of the test situation
(Schiff, 1983; Moore and Frye, 1986) showed that practical problems of the verbal method
are evident as well. Siegal (1991) noted many reasons why a child may misinterpret the
experimenter’s intent. For instance, the child may start doubting because the experimenter asks the same question twice or may consider the task a game. Furthermore, the lack of consensus about which justifications of children demonstrate true knowledge (Brainerd, 1973) constitutes an important problem of a verbal procedure.

Nonverbal versions of Piagetian tasks do not have these theoretical and practical disadvantages. Moreover, the advantages of a nonverbal test are innumerable: more items can be assessed in the same amount of time and other media, like computers and paper-and-pencil tests, are applicable. Training of experimenters and complex coding of tape-recorded justifications are not required. As will be shown below, the problem of false positive and negative errors can be ruled out to a large degree when an adequate methodology is used.

Siegler (1981) designed the Rule Assessment Methodology (RAM) to assess children’s performance on the balance scale task in a nonverbal way. RAM involves a careful selection of item types that elicit specific response patterns. These patterns can be linked to rules, comparable to the original stages. Each rule, represented in Figure 2, is hypothesized to consist of consecutively executed steps. Each step is indicated with a parameter (in italics) that indicates the duration of executing the step. We hypothesize that the time involved with solving a balance scale item equals the sum of the duration of the steps that are completed. This hypothesis leads to a number of predictions, which are tested in a later section.

Complexity of rules increases with development as each rule consists of the steps of the preceding rule, extended with one or more extra steps. Rule I is the simplest rule as it involves only one step, which consists of comparing the numbers of weights ($w$). Participants who use Rule I decide that the scale will tip to the side with the largest number of weights when the numbers are unequal and that the scale will remain in balance when the numbers are equal. Participants who use Rule II also compare the numbers of weights on all items and decide that the scale will tip to the side with the larger number of weights when the numbers...
differ. However, when the numbers are equal, they also compare the distances at which the weights are placed ($d$). Rule III is more complex than Rule II because it contains two additional steps. Rule III-users derive the correct response on simple items by comparing the numbers of weights in the first step and comparing the distances at which the weights are placed in the second step. The first additional step is performed if both the weights and the distances are unequal and includes determining whether the dimensions agree ($a$; i.e., whether the greater weight is on the same side as the greater distance). If the dimensions conflict, the second additional step is performed. It implies “muddling through” or guessing ($g$). Although Rule IV contains the same number of steps as Rule III, Rule IV is more complex because of the complexity of the last step. It includes executing the torque-rule on items with conflicting dimensions ($p$).
Figure 2: Rules defined by Siegler (1981). Also depicted is the compensation-rule.

Lowercase italics $w$, $d$, $a$, $c$, $g$, and $p$ indicate duration of a step. Adapted from van der Maas & Jansen (2003).

To assess the use of these rules, Siegler designed (1976, 1981) six item types. On simple-balance items (“sb”), both arms of the scale hold the same number of weights, equidistant from the fulcrum. On simple-weight items (“sw”), the arms contain unequal numbers of weights, equidistant from the fulcrum. Simple-distance items (“sd”) involve equal numbers of weights, placed at different distances from the fulcrum. On conflict items, one arm contains a greater number of weights, whereas the weights on the other arm are placed at a greater distance. The scale tips to the side with the larger number of weights on conflict-weight items (“cw”), tips to the side with the weights placed at the greater distance on conflict-distance items (“cd”) and remains in balance on conflict-balance items (“cb”). Table 1 shows that each rule results in a specific response pattern on these item types. The indices in italics represent the duration of executing a rule, given item type.
Table 1. Predicted proportion of correct items and RT for each rule, given item type

<table>
<thead>
<tr>
<th>Item type</th>
<th>Example</th>
<th>Rule I</th>
<th>Rule II</th>
<th>Rule III</th>
<th>Compensation</th>
<th>Rule IV</th>
</tr>
</thead>
<tbody>
<tr>
<td>simple-balance</td>
<td><img src="image" alt="Balance Scale" /></td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>w</td>
<td>w+d</td>
<td>w+d</td>
<td>w+d</td>
<td>w+d</td>
</tr>
<tr>
<td>simple-weight</td>
<td><img src="image" alt="Weight Scale" /></td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>w</td>
<td>w+d</td>
<td>w+d</td>
<td>w+d</td>
<td>w+d</td>
</tr>
<tr>
<td>simple-distance</td>
<td><img src="image" alt="Distance Scale" /></td>
<td>.00 a</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>w</td>
<td>w+d</td>
<td>w+d</td>
<td>w+d</td>
<td>w+d</td>
</tr>
<tr>
<td>conflict-balance</td>
<td><img src="image" alt="Conflict Scale" /></td>
<td>.00 b</td>
<td>.00 b</td>
<td>.33 c</td>
<td>-- d</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>w</td>
<td>w</td>
<td>w+d+a+g</td>
<td>w+d+a+c</td>
<td>w+d+a+p</td>
</tr>
<tr>
<td>conflict-weight</td>
<td><img src="image" alt="Conflict Weight Scale" /></td>
<td>1.00</td>
<td>1.00</td>
<td>.33 c</td>
<td>-- d</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>w</td>
<td>w</td>
<td>w+d+a+g</td>
<td>w+d+a+c</td>
<td>w+d+a+p</td>
</tr>
<tr>
<td>conflict-distance</td>
<td><img src="image" alt="Conflict Distance Scale" /></td>
<td>.00 b</td>
<td>.00 b</td>
<td>.33 c</td>
<td>-- d</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>w</td>
<td>w</td>
<td>w+d+a+g</td>
<td>w+d+a+c</td>
<td>w+d+a+p</td>
</tr>
</tbody>
</table>

Note. a Answers that the scale will remain in balance. b Answers that the scale will tip to the side with more weights. c Guesses or muddles through. d Response depends on configuration of weights and distances. w, weight comparison; d, distance comparison; a, deciding whether the weight and distance dimension agree; g, guess or muddle through; c, compensation; p, compare products.

RAM inspired many researchers to study children’s behavior on the balance scale task (Roth, 1991). Ferretti and Butterfield (1989) focused on different populations whereas Halford, Andrews, Dalton, Boag, and Zielinski (2002) and Richards and Siegler (1981) focused on younger children. Others designed alternative methods (Chletsos, De Lisi, Turner...
& McGillicuddy-De Lisi, 1989; Kliman, 1987; Marini & Case, 1994; McFadden, Dufresne, & Kobasigawa, 1987). Siegler and Chen (1998) performed a microgenetic study. Also the effects of training (Ferretti & Butterfield, 1992; Phelps & Damon, 1989; Warton & Bussey, 1988) and the relationship between balance scale performance and other abilities or personality traits (Ferretti, Butterfield, Cah, & Kerkman, 1985; Pauen & Wilkening, 1997; Surber & Gzesh, 1984) were studied. Finally, RAM motivated researchers to model children’s behavior by means of AI models and neural networks (McClelland & Jenkins, 1991; Schmidt & Ling, 1996; Shultz, Mareshal, Schmidt, 1994; van Rijn, van Someren, & van der Maas, 2003).

Some studies resulted in the proposal of alternative rules (Wilkening & Anderson, 1982). The addition-rule (Ferretti, et al, 1985; Normandeau, Larivée, Roulin & Longeot, 1989; Jansen & van der Maas, 1997, 2002) is most frequently discussed. It results in the same response pattern as the buggy-rule (van Maanen, Been & Sijtsma, 1989). On conflict items, the first three steps in these rules are similar to the steps in Rules III and IV: comparing both the values on the weight and the values on the distance dimension and deciding whether the dimensions conflict. At the fourth step, the addition-rule involves adding weight and distance on each side of the scale and comparing the sums, whereas the buggy-rule involves shifting the larger number of weights away from the fulcrum. One weight is removed for every shift, until the distances or the weights of both sides are equal. Halford et al (2002) recognized the compensational character of both rules and we adopt their designation of these rules as “compensation-rule”. Figure 2 contains the representation of the compensation-rule, whereas Table 1 shows that the rule results in the correct response to all simple items, and to some conflict items.

STATISTICAL TEST OF THE RULE ASSESSMENT METHODOLOGY
Although RAM proved to be a useful, nonverbal, method for studying children’s knowledge, the method also met with criticism. Most importantly, the existence of rules was doubted. McClelland (1995) doubted whether humans use explicit rules, because studies by McClelland (1989) and McClelland and Jenkins (1991) showed that even connectionistic models, that do not contain explicit rules, do show rule-like behavior. In these studies, data, generated with simple connectionistic models, were analyzed with RAM. The results lead to the claim that the simple connectionistic models showed rule-governed behavior and that they even acquired knowledge in a stage wise process of rule acquisitions. Ferretti and Butterfield (1986) observed that the classifications in rules varied with the problem set used: Children were classified as users of more complex rules if the items in a set contained larger product differences, than when items with smaller product differences were used. In other words, the items of one type were not homogeneous and the rule classification depended on the items used.

To decide whether children truly use rules, we need empirical criteria for the concept “rule”. Reese (1989) distinguished six criteria for rule inference, of which an important one is that observed behavior is consistent with expected behavior. Although RAM seems an elegant method to assess this criterion, there is a risk that the rules are an artifact of RAM (see Strauss & Levin, 1981). First, RAM does not include any statistical fit measures to evaluate the fit of the rule classification to the empirical data. Second, the criterion of matching responses that is used to classify response patterns (e.g., 20 out of 24 responses correspond to a rule) is chosen arbitrarily and does not lend itself to statistical testing. As the criterion changes with the number of items, comparison of classifications based on different data sets, obtained with different tests, is difficult. Finally, with RAM, only the rules that are known a priori can be detected. A statistical approach to the classification of response patterns to rules is required to solve these problems.
The unobserved rule that a child uses can be regarded as a latent property that determines the child’s responses to a set of balance scale items. The idea of a latent property corresponds to the concept of a latent variable, as postulated in latent structure models (Clogg, 1995). Latent structure models are “models that introduce latent variables to account for the observed pattern of association between the manifest variables” (Lazarsfeld & Henry, 1967, in Heinen, 1996, p. 3). This means that the unobserved latent property (proportional reasoning) determines the behavior of a person’s manifest indicators (balance scale problems).

The latent class model is a latent structure model that deals with categorical latent and manifest variables. It is obvious that the manifest variables are measured on a categorical scale, as the responses to balance scale problems are “left side down”, “right side down”, and “the two sides balance”. It is assumed that the latent variable, the ability to solve balance scale items, is also only measurable on a categorical scale, as a child’s ability to solve balance scale problems is assumed to correspond to mastery of either Rule I, Rule II, Rule III, or Rule IV. These rules can not be viewed as points on a continuous scale as the procedures that are associated with the rules differ qualitatively from each other. The rules are increasingly complex but cannot easily be ordered. Consider for example the expected number of correct responses given Rule II and Rule III. Children who use Rule II answer more conflict-weight items correctly, whereas children who use Rule III answer more of the other conflict items correctly (see Table 1). However, the expected numbers of correct items are equal. Hence, it can be assumed that the measurement level of the latent ability of proportional reasoning is categorical and that each rule forms a category.

Collins and Wugalter (1992) and Rindskopf (1987) already demonstrated the usefulness of LCA for the analysis of developmental data. With LCA, the number of rules, and their response patterns, can be determined. In these response patterns, children’s error processes
can be modeled. A child’s response may be inconsistent with the rule that the child is using because of carelessness, for instance (Rindskopf, 1987). The deviation from the expected response pattern can be accommodated in a latent class model. Hence, the criterion used to classify children into classes (i.e., to rules) is based on a statistical criterion, rather than an arbitrary criterion. LCA also eases the comparison of classifications based on different data sets, collected with different tests. Another advantage is that LCA does not require information on the content of rules but detects clusters of response patterns in the data, which can later be interpreted as (alternative) rules. Most importantly, LCA can falsify the hypothesis concerning rule use, because a latent class model, associated with rule use, can be tested statistically.

**Latent class analysis**

LCA divides the population in a finite number of latent classes. Within each class, the manifest variables are assumed to be statistically independent. A latent class model consists of unconditional probabilities, representing the proportions of the classes, and conditional probabilities, representing the probabilities of giving the response “left side down”, “right side down”, or “balance” on a particular balance scale item, within a given latent class.

We test whether two items are of the same difficulty, for subjects in a certain latent class, by restricting the conditional probabilities of these items, in the given class, to be equal to each other. In this way, Ferretti and Butterfield’s (1986) claim that rule classification depends on the product difference of the items in a specific set, can be tested. The claim is falsified when the conditional probabilities, associated with items of one type, can be restricted to be equal across items, within each latent class.

To select a model, we first determine the fit of a two-class model by considering the likelihood ratio, $G^2$, with respect to the number of degrees of freedom ($df$). In addition to the
theoretical distribution of $G^2$, we use an empirical distribution, which is obtained by means of parametric bootstrapping (Langeheine, Pannekoek, & Van de Pol, 1995). We increase the number of classes until a model is found with an insignificant $G^2$. Next, restrictions are applied to the unrestricted model and the tenability of the restrictions is judged by applying the likelihood-ratio difference test. Additionally, models with insignificant $G^2$ are compared by means of the Bayesian Information Criterion (BIC), which combines fit and parsimony. The model with the lowest value is selected. Subjects are assigned to the latent classes by means of posterior probabilities, based on the selected model.

We apply LCA to two empirical data sets, and to a data set that was derived with the PDP model for the balance scale task (McClelland, 1989). This allows for testing the claim that the PDP model can simulate children’s behavior on the task, which was supported by the results of the RAM. Analyses are performed with PanMark (van de Pol, Langeheine, & de Jong, 1996).

Application of LCA to responses to balance scale items

*Data sets IA and IB* The first empirical data set (Data set IA), collected by van Maanen et al. (1989), consists of the responses of 484 children to a paper-and-pencil test of 25 balance scale items. Children were in grades seven and eight. The simulated data set (Data set IB) was generated with the same items and consists of the same number of response patterns. Both data sets consist of dichotomous responses (false/correct). Analyses were carried out for each item type individually to test the hypothesis that children respond consistently to items of a single type. As Table 1 suggests, the probability of answering an item correctly depends only on the item type (and the rule that a child applies). Any other item characteristics (e.g., product difference) are not expected to affect response probabilities. Equality restrictions between items within each class, are introduced to test consistency.
As an example, we present the results of the analyses of the cb-items. LCA of responses to this item type should result in three latent classes. One class is expected to consist of children who answer cb-items correctly because they use Rule IV or the compensation-rule. The compensation-rule does not necessarily result in a correct response to cb-items, but it does given the items in this test. A second class is expected to consist of children who use Rule III and who guess on cb-items. A third class is expected to consist of children using Rule I or Rule II, who will answer that the scale will tip to the side with the larger number of weights (see Table 1) and hence fail cb-items. Table 2 shows the fit measures of the latent class models of both data sets.
Table 2. Fit-measures of latent class analyses of conflict-balance items, for data-sets IA and IB.

<table>
<thead>
<tr>
<th>Model</th>
<th>$G^2$</th>
<th>df</th>
<th>$p(G^2)$</th>
<th>pb($G^2$)</th>
<th>$\Delta G^2$</th>
<th>$\Delta df$</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data set IA (empirical data set)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>84.77</td>
<td>20</td>
<td>.00</td>
<td>.00</td>
<td>-</td>
<td>-</td>
<td>152.54</td>
</tr>
<tr>
<td>3</td>
<td>33.16</td>
<td>16</td>
<td>.01</td>
<td>.01</td>
<td>-</td>
<td>-</td>
<td>125.58</td>
</tr>
<tr>
<td>4</td>
<td>19.80</td>
<td>11</td>
<td>.05</td>
<td>.07</td>
<td>-</td>
<td>-</td>
<td>143.02</td>
</tr>
<tr>
<td>5</td>
<td>13.30</td>
<td>8</td>
<td>.00</td>
<td>.06</td>
<td>-</td>
<td>-</td>
<td>155.01</td>
</tr>
<tr>
<td>4e</td>
<td>34.19</td>
<td>24</td>
<td>.08</td>
<td>.09</td>
<td>14.39</td>
<td>13</td>
<td>77.32</td>
</tr>
<tr>
<td>Data set IB (PDP data set)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>134.47</td>
<td>21</td>
<td>.00</td>
<td>.00</td>
<td>-</td>
<td>-</td>
<td>196.04</td>
</tr>
<tr>
<td>3</td>
<td>68.09</td>
<td>17</td>
<td>.00</td>
<td>.00</td>
<td>-</td>
<td>-</td>
<td>154.29</td>
</tr>
<tr>
<td>4</td>
<td>27.06</td>
<td>14</td>
<td>.00</td>
<td>.00</td>
<td>-</td>
<td>-</td>
<td>138.17</td>
</tr>
<tr>
<td>5</td>
<td>13.53</td>
<td>11</td>
<td>.13</td>
<td>.13</td>
<td>-</td>
<td>-</td>
<td>138.07</td>
</tr>
</tbody>
</table>

*Note.* Adapted from Jansen & van der Maas (1997). pb, bootstrapped p-value.

In case of the empirical data set, a four-class model is chosen, as this is the model with the smallest number of latent classes that fits the data. In model 4e, the conditional probabilities were restricted to be equal within each of the first three classes. When all five conditional probabilities of the fourth latent class were constrained to be equal, the model was rejected. Inspection of the item characteristics revealed that all cb-items could be solved by performing the buggy-rule. However, only one buggy was needed to solve the third and fourth item, whereas two buggies were needed to solve the remaining items. Hence, we restricted the conditional probabilities of the first, second, and fifth item to the same value and restricted the conditional probabilities of the third and the fourth item to the same value.
The likelihood-difference test showed that the restrictions of model 4e were acceptable \((p > .05)\). The BIC value of model 4e was also lower than that of the unrestricted 4-class model.

The selected model is presented in Figure 3. The largest class \((p = .36)\) showed low conditional probabilities of answering cb-items correctly, which matched the expected responses of children who use either Rule I or Rule II. The second class \((p = .26)\) showed conditional probabilities that approximated a guessing level (Rule III). The third class \((p = .25)\) showed high conditional probabilities of answering cb-items correctly, and was expected to comprise children who used Rule IV or the compensation-rule. The smallest class \((p = .13)\) was unexpected. It was characterized by high probabilities of answering cb-items 3 and 4 correctly, but probabilities that suggested guessing on the remaining items. Apparently, children in this class used the buggy-rule but applied it poorly. They seemed to guess when the conflict item was not reduced to a simple item after making one buggy.

In case of the data set generated with the PDP model, it was more difficult to find a model that described the data adequately. The fit measures in Table 2 indicate that five latent classes were needed to describe the data, but the five-class model, represented in Figure 3, was highly sensitive to the choice of starting values. This instability suggests that the LCM is not suitable. Moreover, only a single class was compatible with Siegler’s rules (showing a response pattern that was characterized by low conditional probabilities of answering cb-items correctly and hence matched Rule I and Rule II). This class accounted for .38 of the cases.
Figure 3: Latent class models of children's data and data generated with the PDP model (NN) of McClelland (1989). Sizes of the triangles indicate proportions of the classes. The items are five conflict-balance items. Adapted from Jansen & van der Maas (1997).

The analyses of the cb-items were representative of the analyses of the remaining item types. The latent classes in the models of the empirical data set matched Siegler’s rules (Rule I, Rule II) or could be interpreted as stemming from plausible alternative rules (compensation-rule). The evidence for Rule IV was less convincing, but the children in data set IA were probably too young to display this rule. Rule III was detected in some, but not all latent class models. Simulation studies showed that it is difficult to detect a guessing rule like Rule III with LCA. Some children, who used a buggy version of the compensation-rule, had difficulty applying this rule as they resorted to guessing when it took more than one buggy to
reduce the conflict item to a simple item. Equality restrictions showed that children responded consistently to items of the same type. This result conflicts with the inconsistency Ferretti and Butterfield (1986) observed.

Data set II The second empirical data set (Data set II) features in Jansen and van der Maas (2002), and consists of the responses of 805 subjects to another test of 25 balance scale items. Subjects’ ages ranged from 5 to 19 years. Data set II consists of trichotomous responses (left/balance/right). Analyses of individual item types showed that responses to items of one type were homogeneous. It then can be concluded that an item is representative for other items of the type. We present a latent class model of responses to a combination of item types. Only with a combination of item types can all Siegler’s rules be detected. We selected the responses to two sd-items, two cw-items and two cb-items. The items that were used to collect data set II were constructed in such a way that the use of the compensation-rule resulted in the incorrect response “in balance” to cw-items, and in the correct response to the remaining conflict items. The expected response patterns for Siegler’s rules and the compensation-rule differed importantly in this set of items (see Table 3). We expected a five-class model, with each class representing a rule.
Table 3. Expected latent classes for latent class model of a combination of items

<table>
<thead>
<tr>
<th>Conditional probabilities</th>
<th>distance</th>
<th>conflict-weight</th>
<th>conflict-balance</th>
<th>interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>left</td>
<td>balance</td>
<td>right</td>
<td></td>
</tr>
<tr>
<td></td>
<td>left</td>
<td>balance</td>
<td>right</td>
<td></td>
</tr>
<tr>
<td></td>
<td>left</td>
<td>balance</td>
<td>right</td>
<td></td>
</tr>
<tr>
<td></td>
<td>left</td>
<td>balance</td>
<td>right</td>
<td></td>
</tr>
</tbody>
</table>
| Note. The probabilities for the correct response are printed in italics. For the conflict-balance items, the right side has the larger number of weights. Adapted from Jansen & van der Maas (2002).

A seven-class model was selected as this was the first model that showed an adequate fit to the data ($G^2 (671) = 174.70$, $pb(G^2) > .05$) and because this model had a lower value of BIC than models with more classes. Table 4 shows the selected model. Table 4 only contains the conditional probabilities of one item of each type, to save space, and because the probabilities were quite similar.
Table 4. Latent class model of responses to a combination of items

<table>
<thead>
<tr>
<th></th>
<th>distance 4</th>
<th></th>
<th>conflict-weight 4</th>
<th></th>
<th>conflict-balance 4</th>
<th></th>
<th>interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$p(t)$</td>
<td>left</td>
<td>balance</td>
<td>right</td>
<td>left</td>
<td>balance</td>
<td>right</td>
</tr>
<tr>
<td>1</td>
<td>.27</td>
<td>.00</td>
<td>.96</td>
<td>.04</td>
<td>.96</td>
<td>.01</td>
<td>.02</td>
</tr>
<tr>
<td>2</td>
<td>.15</td>
<td>.70</td>
<td>.21</td>
<td>.09</td>
<td>.87</td>
<td>.13</td>
<td>.00</td>
</tr>
<tr>
<td>3</td>
<td>.09</td>
<td>.98</td>
<td>.02</td>
<td>.00</td>
<td>.44</td>
<td>.23</td>
<td>.34</td>
</tr>
<tr>
<td>4</td>
<td>.11</td>
<td>1.00</td>
<td>.00</td>
<td>.00</td>
<td>.82</td>
<td>.11</td>
<td>.07</td>
</tr>
<tr>
<td>5</td>
<td>.18</td>
<td>.97</td>
<td>.01</td>
<td>.02</td>
<td>.03</td>
<td>.95</td>
<td>.02</td>
</tr>
<tr>
<td>6</td>
<td>.02</td>
<td>.17</td>
<td>.17</td>
<td>.67</td>
<td>.92</td>
<td>.08</td>
<td>.00</td>
</tr>
<tr>
<td>7</td>
<td>.17</td>
<td>.93</td>
<td>.07</td>
<td>.00</td>
<td>.24</td>
<td>.66</td>
<td>.10</td>
</tr>
</tbody>
</table>

*Note. $t =$ latent class, $p(t) =$ proportion of latent class $t$. The probabilities for the correct response are printed in italics. For each type, item 4 is shown. For the conflict-balance items, the right side has the larger number of weights. Rule III/comp is a combination of Rule III and the compensation-rule. SDD is the Smallest Distance Down Rule. Adapted from Jansen & van der Maas (2002).*

Comparison of Tables 3 and 4 shows that the first five classes in the selected model matched the five expected latent classes. The children in the first additional class ($p = .02$) responded that the scale would tip to the side with the smallest distance on sd-items. On conflict items, these children also responded that the scale would tip to the side with the smallest distance (or the largest number of weights). These children perhaps thought that the scale always tipped to the side with the smallest distance (SDD-rule). The responses of the children in the second additional class ($p = .17$) to the sd-items and to the cw-items were
similar to those of the children in the fifth class, but the responses to the cb-items differed. Possibly, these children used some mix of Rule III and the compensation-rule.

**Conclusion** The results of LCA suggest that children actively employ rules in solving balance scale problems. The value of the application of LCA to the balance scale task has been confirmed by the latent class analyses of balance scale data reported by Boom, Hoijting, and Kunnen (2001). They demonstrated that even LCA of large data sets, with 20 items, is feasible. An additional advantage of the application of LCA is that we were able to detect alternative rules, like the compensation-rule. Contrary to the rule classifications of RAM, the latent classes reflect the structure in the observed data and arise independently from the rules postulated in a theory.

The difficulties of interpreting the LCA models for the PDP data set were striking. Although fitting latent class models was possible, the results were often unstable (i.e., dependent on starting values). Furthermore, the response patterns of the classes did not show consistency within item types, and hardly matched Siegler’s rules or any plausible alternative rule. Probably, no rules underlie the response patterns of this connectionist model, in spite of the results that were obtained by applying RAM (McClelland, 1989; McClelland & Jenkins, 1991). These results are in accordance with the results of Raijmakers, van Koten, and Molenaar (1996). It seems that the RAM is too liberal and may falsely suggest the presence of rules. Strauss and Levin's (1981) criticism that rules are an artifact of the methodology seems correct. However, application of the statistical technique of LCA allows for the falsification of the rule theory. We conclude that it is possible to reveal the rules children use by combining RAM with LCA, even without asking children to clarify their responses.

The restrictions that were applied to items of the same type proved that items of one type are homogeneous. The influence of product difference on rule classification that Ferretti and
Butterfield (1986) detected, was not observed here. A reason for this may be that we did not increase the product differences between the two sides of the scale like Ferretti and Butterfield did. As claimed elsewhere (Jansen & van der Maas, 1997, p. 327), this indicates that their conclusion is based only on the extreme values of the product differences used in their experiment. As the conditional probabilities associated with the items can be restricted to be equal for each item type, it is possible to analyze the sum scores of each item type with finite mixture analysis. As Turner and Thomas (2002) noted, this technique can deal with smaller samples and a larger number of variables. However, the analysis of cb-items showed that items of a type are not necessarily homogeneous for all rules.

RULES AND RESPONSE TIMES

We have proposed a new and promising method of validation of the rules in van der Maas and Jansen (2003) by using RTs. In this section we discuss the main results of this study. Figure 2 and Table 1 represent the predictions concerning RTs that can be derived from Siegler’s (1981) rule theory. The predictions concerning RTs are tested by fitting regression models to RT data. Next, we propose and test several new predictions.

Table 1 shows that the RTs on all item types are predicted to equal $w$ for participants who use Rule I. The RTs for Rule II-users on items with different numbers of weights, i.e., sw- and conflict items, are predicted to equal $w$, whereas the RTs on items with equal numbers of weights, i.e., sb- and sd-items, are expected to be longer, i.e., $w + d$. The RTs of participants who use Rule III, Rule IV, or the compensation-rule are expected to equal $w + d$ on simple items. On conflict items, the RTs are predicted to equal $w + d + a + g$, $w + d + a + p$, or $w + d + a + c$ for users of Rule III, Rule IV, or the compensation-rule, respectively. Of course, the RTs for all rules also include a constant amount of time for common processes like the motor response. However, the duration of this constant cannot be estimated.
independently of \( w \) as both processes always occur together.

Testing the RT predictions of the basic model

The parameters \( w, d, a, g \) and \( c \), derived from Siegler’s (1981) basic model, can be used as independent variables in regression analysis in which RT is the dependent variable. To perform this type of regression analysis, we collected the responses (left, balance, right) and the RTs of 147 children (6 to 15 years of age) and 44 undergraduate psychology students for 70 balance scale items, presented in a computer test. The item set consisted of 10 sets of 7 items (sb, sw, sd, cbb, cw, cd, cba). Both cba and cbb items are conflict-balance-items, but cba items can be solved correctly with the compensation-rule whereas cbb items can not.

Rule assignment was done with RAM and with iterative cluster analysis on the sum scores per item type and not with latent class analysis in view of the large number of items (70). The results of these methods agreed well (Cohen’s Kappa of .85). The classification that resulted from the cluster analysis was used in the RT analyses.

RT data were analyzed per subject and per item, allowing for the incorporation of subject characteristics (like age) and quantitative item characteristics (like product difference). Model A contained parameters \( w, d, g, c, \) and \( p \), but not parameter \( a \) because of identification problems. Model A also contained parameters that modeled the “law of practice” \( (i e^{-bI}; \) where \( I \) is the order of the item in the test). The law dictates that the speed of responding increases during administration of a test (see, e.g., Thorndike, 1913). In model A, comparing weights \( (w) \) was estimated 2.75 s, whereas comparing distances \( (d) \) was estimated at –0.22 s. The duration of guessing, performing the production rule, and performing a compensation process were estimated at 2.28 s \( (g) \), 2.62 s \( (p) \), and 2.39 s \( (c) \), respectively. The law of practice was modeled as \( 1.83e^{-0.07I} \).
Extending the basic model. Model A is extended with additional parameters, resulting in Model B. Different parameters were introduced and estimated for comparing equal and comparing unequal numbers of weights. Furthermore, the basic process of comparing equal weights (estimated at 2.02 s) was slowed with 0.18 s for each extra weight in each pile. Comparing unequal weights took 2.60 s, independent of the size of the difference between the numbers of weight.

From model B, it was concluded that Rule II-users compared distances at any item, whether the numbers of weight were equal or not. Although the response patterns showed that Rule II-users always answered that the scale would tip to the side with the larger number of weights, their RTs showed that they did consider the distance dimension when the
numbers of weights differed.

Deciding that the distances were identical was easy when the piles were placed at the 
pegs next to the fulcrum or at the furthest pegs from the fulcrum. The decision took 0.66 s 
more when the piles were placed near the center of each arm. Deciding that the piles of 
weight were placed at different distances was mediated by the difference between the 
distance on the left and the distance on the right side. An increase of distance difference of 
one, slowed users of Rule II with 0.34 s.

Users of the complex rules, Rule III, the compensation-rule, and Rule IV, were 
considered to compare weights and to compare distances at each balance scale item. The 
steps proceeded in the same way as for users of Rule I and Rule II. However, the process of 
comparing unequal distances turned out to be different for users of complex rules. An 
increase of distance difference of one shortened RT with 0.41 s. Probably, the difference 
between the distances was more easily noted when the difference was large. Parameter $a$ was 
estimated at 1.29 s.

It was concluded that the RT of Rule III-users was 0.27 s longer when the sum of the 
difference between the values on the weight dimension and the difference between the values 
on the distance dimension increased with one. Possibly, these participants could not decide 
when both dimensions were salient. It seems likely that their decision is based on the careful 
consideration of the possibilities and does not imply pure guessing as the RTs were relatively 
long and comparable to those associated with the compensation-rule and Rule IV.

Although the response patterns for the addition-rule and the buggy-rule do not differ, the 
predictions concerning the RTs associated with the rules do differ. We conclude that 
participants used a buggy-rule and not an addition-rule because model B demonstrated that 
RT was affected only by parameters that were related to predictions that followed from the 
buggy-rule: RT increased with 1.13 s for each required buggy and increased with 0.23 s for
each weight in the pile that needed to be shifted.

Model B showed that Rule IV-users consider whether a conflict item contains any symmetries: RT was 0.37 s shorter when the distance, at which the weights were placed on one arm, was equal to the number of weights, on the other arm (see Figure 1 for an example). Moreover, RT was lengthened with 0.12 s when the sum of products of weight and distance increased with one, resulting in an increase of product difficulty.

An extra parameter was added to improve the description of the RTs of users of Rule IV. This parameter referred to the hesitation, lasting 0.74 s, to answer that the scale would tip to the side with the larger number of weights. Although the parameter did improve the fit of the model, we do not consider it a very reliable parameter.

The base speed rate of older children was faster than that of younger children. Finally, a positive relation was observed between rule inconsistency (deviance of the observed response pattern from the cluster pattern) and increase of RT.

Model B described the data significantly better than model A, which demonstrated that inclusion of quantitative item characteristics improves the prediction of RTs considerably.

**Weight-distance items** The last 6 items of the balance scale test were so-called weight-distance items. Weight-distance items are items in which the larger number of weights is on the side of the larger distance. All subjects should solve these simple items easily. Users of Rule I and II are expected to take into consideration the weight dimension \( w \) only, whereas users of higher rules are expected to look at the distance dimension \( d \) too, and even check whether the larger weight is on the side with the larger distance \( a \). This results in the provoking prediction that older and more advanced subjects are expected to respond more slowly to these easy items. This prediction was confirmed by the data. Mean RT for Rule I
and II was 2.63 s, whereas the mean RT for Rule III, the compensation-rule and Rule IV was 3.96 s. This large difference was significant ($t (52) = -5.02, p < .01$).

Conclusions

The analyses of the RTs validated Siegler’s rule models to a large extent. Next, it was possible to refine the rule models greatly by using both subject (age and rule inconsistency) and item characteristics (e.g., number of buggies). It was concluded that the compensation-rule is actually a buggy-rule and not an additive rule and that Rule II-users always consider the distance dimension but that they nevertheless neglect it in their response. This result matches Siegler’s (1981) observation that many children, who demonstrated Rule II in their response patterns, did indicate the use of the distance dimension in their explanation afterwards. Further study of this finding is required. Finally, the results showed that inconsistent rule use was associated with increased RT. A complete analysis of these data can be found in van der Maas and Jansen (2003).

TRANSITION FROM RULE I TO RULE II

Having established a clear idea of children’s strategies for solving balance scale problems, it is time to study whether the transitions between the strategies proceed either continuously or discontinuously. Although the character of transitions between stages constitutes a central and recurrent issue in developmental psychology, criteria to distinguish the two types of development were lacking (Brainerd, 1978). Recently, van der Maas and Molenaar (1992, 1996) proposed to use catastrophe theory to test for discontinuities. Catastrophe theory (Thom, 1975) is a general mathematical theory of transitions, which are defined as large sudden jumps as function of small continuous changes in independent variables. The most popular model of catastrophe theory is the so-called cusp model which has two independent variables. For instance, the freezing and melting of water and ice can be
modeled with the cusp model. This transition is controlled by two independent variables, temperature and pressure.

Cusp-like processes are characterized by a number of phenomena, so-called catastrophe flags, among which are the sudden jump, bimodality, inaccessible region, critical slowing down, hysteresis and divergence (Gilmore, 1981). The last two are convincing indicators of discontinuity. Hysteresis occurs when the sudden jump position depends on the direction of change in the normal variable. For instance, ice melts at 0°C but water freezes (in disturbance free conditions) at –4°C. Divergence means that the system is forced to choose between two alternative modes when the splitting variable increases. For instance, at low pressure, water at –2°C can be in an in-between state. When pressure increases to 1 atmosphere, either the solid state or the fluid state is selected and the in-between state becomes unstable. Figure 5 depicts the cusp model, where $\alpha$ is the normal variable and $\beta$ is the splitting variable, and the main catastrophe flags. The proposal of van der Maas and Molenaar (1992) is to use the flags as criteria for detecting developmental transitions.
A cusp model for the Rule I to Rule II transition

Bimodality is a necessary, although not sufficient, indicator of transitions. Bimodality is clearly shown in the distribution of sum scores of sets of distance items. The distribution shows a strong distinction between two modes of responses: all incorrect (Rule I) or all correct (Rule II or a more complex rule) (Raijmakers, van Koten, & Molenaar, 1996; Jansen & van der Maas, 1997, 2001; see also Section 2). This inspired us to think about a cusp model.

Figure 5: The cusp model of the Rule I to Rule II transition. Adapted from Jansen & van der Maas (2001).
model of the Rule I to Rule II transition. A good choice of the dependent variable is the percentage correct (or some transformation of this percentage) on sets of distance items as Rules I and II differ only in the responses to these items. A selection of independent variables, i.e., the normal and the splitting variable, is more difficult. The normal variable gives rise to sudden jumps and hysteresis, the splitting variable to divergence.

According to Siegler and Chen (1998), the main difference between Rules I and II is the ability to encode distance. We emphasized the saliency of the distance dimension to increase the awareness of the distance dimension of Rule I-users. Increase of awareness may result in encoding the dimension and, eventually, in a switch to using Rule II. Saliency was emphasized by increasing the distance difference between the two sides of the scale in distance items. Hence, we chose the difference in distance between the weights on the left and the distance of the weights on the right side of the scale as the normal variable.

We chose the number of weights used in a distance item as splitting variable. This choice is based on the premise that more weights increase the confidence of both Rule I- and Rule II-users. Since Rule I-users only take the number of weights into account, a more salient weight dimension should increase the probability that they answer according to this rule. On the other hand, increase of weights increases the product difference between left and right. Rule II-users may have more confidence in their answer because the momentum for the side with the larger distance increases. This polarization for items with equal distance differences and different amounts of weights is a divergence effect (see Figure 5).

With the choice of these two variables, the cusp model is complete, as depicted in Figure 5. The following main predictions can be derived from this model:

1) Sudden jump: by increasing the distance difference, some (transitional) Rule I-users spontaneously jump to Rule II.
2) Bimodality/inaccessibility: scores on a set of distance items are bimodally distributed, the intermediate scores have low probability.

3) Divergence: When the number of weights increases, the distribution of the observed scores becomes increasingly bimodal.

4) Hysteresis: the distance difference at which Rule I-users jump to Rule II when the distance difference is increased, is larger than the distance difference at which Rule II-users switch to Rule I, when the distance difference is decreased.

5) Critical slowing down: subjects are easily perturbed in the area where the transition is likely to occur (between the bifurcation lines, see Figure 5). A possible effect is an increase in RTs (van der Maas, Raijmakers, Hartelman, & Molenaar, 1998).

Divergence and Hysteresis

The main difference between catastrophe theory and other models for developmental transitions (Brainerd, 1979; Eckstein, 2000; Thomas & Lohaus, 1993; van Geert, 1998; Wilson, 1989) is that catastrophe theory explains instead of assumes abruptness. In addition, the criteria divergence and hysteresis are unique to catastrophe theory. Hence, in our studies we have focused on these two catastrophe flags.

We used six distance items with a distance difference of two to test for divergence. Three of these items have five weights and three items have a single weight. The bimodality on the first three items should be more pronounced.

We used special series of items to test for hysteresis. In these series, we started with a distance item with a distance difference of one. Distance difference increased to five (the scale in the experiment had six pegs) in the next four items, and decreased again, in the next four items, to one. In a control series, which should not lead to jumps and hysteresis, the grayness of the weights was increased and decreased. The control and hysteresis series were
counterbalanced. In a second experiment we also included a series in which distance
difference was first decreased and then increased.

Given these hysteresis series of nine items, six possible types of patterns can occur, of
which five are predicted by the cusp model. The first two patterns contain either correct
responses only (Rule II) or incorrect responses only (Rule I). The “delay” pattern refers to the
hysteresis pattern: the jump to Rule II takes place at higher distance difference than the jump
back to Rule I. In the “enhanced contrast” pattern the jump to Rule II takes place earlier than
the jump back from Rule II to Rule I. This pattern should not occur according to the cusp
model. In the “Maxwell” pattern, both jumps take place at the same distance difference. The
“sudden jump” pattern implies that subjects switch from Rule I to Rule II but never switch
back, which can be interpreted as a strong delay pattern. Finally, residual irregular patterns
may occur.

Experiments

In our first experimental study, we administered a paper-and-pencil test to 314 children
(between six and ten years old). The test included practice items, a pre- and a post-test to
determine the rule used, and the hysteresis, divergence and control series. The pretest
analysis showed that about 50 % of the children responded according to Rule I, 40 %
responded according to Rule II and about 10 % responded inconsistently. Most of the
children who used Rule I and Rule II were stable rule-users and were not expected to be
sensitive to our manipulations. This was the reason to test a large group of children.

The divergence manipulation proved to be unsuccessful. Both the distribution of scores
for the items with one weight and the distribution of scores for the items with five weights
were strongly bimodal. The distributions were very similar ($\chi^2(3, N = 314) = 1.97, p = .58$).
The hysteresis manipulation was more successful. We classified response patterns with both the RAM and LCA. For the later method, we formulated a confirmatory latent class model that included classes that corresponded to the five expected patterns (Rule I, Rule II, delay, Maxwell, sudden jump) and a residual group. Classes that contained response patterns associated with hysteresis (delay, Maxwell, sudden jump) were necessary to obtain a fitting model, whereas classes that contained response patterns associated with enhanced contrast worsened the model significantly. Table 5 shows the proportions of hysteresis patterns for the hysteresis and the control test, for the analysis with both RAM and LCA.

As expected, most subjects stuck to their rule. About 15% of the subjects (control test: 4%) showed delay, Maxwell, or sudden jump patterns. In a second experiment, we found similar results. At series in which distance difference first decreased (from 5 to 1) and then increased, the percentage of hysteresis patterns was a little lower but still statistically significant.
Table 5. Proportions of hysteresis patterns on hysteresis test and control test

<table>
<thead>
<tr>
<th></th>
<th>LCA</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Hysteresis test</td>
<td>Control test</td>
<td>Hysteresis test</td>
<td>Control test</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rule I</td>
<td>.341</td>
<td>.395</td>
<td>.315</td>
<td>.357</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rule II</td>
<td>.350</td>
<td>.427</td>
<td>.325</td>
<td>.401</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Delay</td>
<td>.029</td>
<td>.003</td>
<td>.022</td>
<td>.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maxwell</td>
<td>.035</td>
<td>.000</td>
<td>.035</td>
<td>.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jump upward</td>
<td>.096</td>
<td>.038</td>
<td>.083</td>
<td>.029</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Residual</td>
<td>.092</td>
<td>.073</td>
<td>.162</td>
<td>.150</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Missing values</td>
<td>.057</td>
<td>.064</td>
<td>.057</td>
<td>.064</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

N = 314. Adapted from Jansen & van der Maas (2001).

Model evaluation

The present results indicate that the transition from Rule I to Rule II shows important characteristics of a genuine phase transition. The evidence for bimodality and inaccessibility is convincing as the score distributions of sets of distance items are strongly bimodal. Although we did not find evidence for divergence, we do not think that the failure of the divergence test falsifies the whole idea of a cusp-like process. Either our choice of the control variable (number of weights) or our operationalization of this variable was not adequate. Moreover, the hypothesis of discontinuity is confirmed by the results supporting the other flags, especially hysteresis. We showed, in various ways, that the small number of subjects showing hysteresis in their responses could not be attributed to chance. In addition, in the analysis of the RTs (van der Maas & Jansen, 2003; see Section 3), we found evidence for an increase in RTs when responses were inconsistent. This effect is particularly strong for Rule
I-users. This phenomenon may be interpreted as an instance of another important catastrophe flag, i.e., "critical slowing down".

A DEVELOPMENTAL MODEL FOR PROBLEM SOLVING BEHAVIOR ON THE BALANCE SCALE TASK

Consistency of responding during the administration of a balance scale test was analyzed to study whether children who use Rule II, Rule III, Rule IV, and the compensation-rule do so as consistently as children who use Rule I. For this analysis, we used empirical data set II (Section 2). For this data set, participants responded to five blocks of items. Responses to the second and third block of the test were analyzed separately from the responses to the last two blocks of the test, with LCA. The results of the first two blocks were similar to those of the last two blocks, which were reported in Section 2. LCA of the data demonstrated evidence for the use of Rule I, Rule II, Rule III, Rule IV, the compensation-rule, a combination of Rule III and the compensation-rule, and the SDD-rule.

Children were assigned to the most probable latent class on both parts of the test by means of posterior probabilities. The results of the assignments are represented in Table 6. Use of latent Markov models is preferred to model rule consistency, but, as Turner and Thomas (2002) noted, using this method causes huge computational problems because of the large number of parameters.
The application of Rule I was found to be stable as almost .86 of the children who used this rule on the first part of the test continued to use it on the later part. This supports our claim that the transition from Rule I to Rule II happens discontinuously. Contrary, only .55 of the children who used Rule II on the first part continued to use this rule on the later part of the test. Many children switched to using either Rule III (.20) or the mix of the compensation-rule and Rule III (.14). The use of the compensation-rule, Rule III, and the combination of Rule III and the compensation-rule was also quite inconsistent. Children, who were at this level of performance, seemed to muddle through the rules in their repertory: Sometimes they compared sums, sometimes they guessed, etc. However, there was a considerable proportion of children (.62) who consistently applied the compensation-rule. Another interesting finding was that children, who used Rule III, or the mix of the

Table 6. Turnover table for rule use on first and last part of the test

<table>
<thead>
<tr>
<th>Last part</th>
<th>Rule I</th>
<th>Rule II</th>
<th>Rule III</th>
<th>compensation</th>
<th>Rule IV</th>
<th>SDD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rule I</td>
<td>.855</td>
<td>.120</td>
<td>.013</td>
<td>.004</td>
<td>.004</td>
<td>.000</td>
</tr>
<tr>
<td>Rule II</td>
<td>.016</td>
<td>.549</td>
<td>.197</td>
<td>.025</td>
<td>.139</td>
<td>.049</td>
</tr>
<tr>
<td>First part</td>
<td>.063</td>
<td>.188</td>
<td>.458</td>
<td>.042</td>
<td>.125</td>
<td>.104</td>
</tr>
<tr>
<td>Rule III</td>
<td>.000</td>
<td>.005</td>
<td>.081</td>
<td>.621</td>
<td>.207</td>
<td>.086</td>
</tr>
<tr>
<td>compensation</td>
<td>.030</td>
<td>.212</td>
<td>.081</td>
<td>.192</td>
<td>.394</td>
<td>.091</td>
</tr>
<tr>
<td>Rule III/comp</td>
<td>.000</td>
<td>.030</td>
<td>.000</td>
<td>.015</td>
<td>.045</td>
<td>.910</td>
</tr>
<tr>
<td>Rule IV</td>
<td>.000</td>
<td>.091</td>
<td>.091</td>
<td>.182</td>
<td>.000</td>
<td>.000</td>
</tr>
<tr>
<td>SDD</td>
<td>.000</td>
<td>.091</td>
<td>.091</td>
<td>.182</td>
<td>.000</td>
<td>.000</td>
</tr>
</tbody>
</table>

Note. Rule III/comp is a combination of Rule III and the compensation-rule. Adapted from Jansen and van der Maas (2002).
compensation-rule and Rule III on the first part of the test, switched to using Rule II on the second part of the test. The high proportion (.91) of children who used Rule IV on both parts of the test, suggested that, once children have learned the correct rule, they always apply it. No conclusions are drawn on the consistency of the use of the SDD-rule as only a few children were identified as users of this rule.

In summary, both the use of Rule I and Rule IV was quite consistent, but there seemed to be much switching between Rule II, Rule III, the compensation-rule, and the combination of Rule III and the compensation-rule. The progress of children who start by using Rule II to using Rule III or the compensation-rule can be explained by spontaneous learning. Children who use Rule II already know that the distance dimension can be important. Their RTs indicate that they consider it at any item (see Section 3), but their responses indicate that they only incorporate it in their strategy when the weights are equal. Merely presenting these children with balance scale items may sensitize them to the distance dimension and may convince them of the importance of the distance dimension. However, they do not know how to combine the two dimensions yet. This is consistent with the interpretation of Rule III. A different mechanism than learning must be responsible for regressing from Rule III or the compensation-rule to Rule II. Finally, many children switched between Rule III, the combination of Rule III and the compensation-rule, and the compensation-rule. This finding supports the hypothesis that children who use Rule III sample from an ensemble of strategies and that switching between these strategies is inherent to Rule III.

Switching between strategies seems to contrast with a so-called staircase model that underlies Siegler’s hierarchy of rules. In a staircase model of development, children apply a rule for a long period and then suddenly shift to a next rule that is associated with a more advanced level of ability. The staircase model implies consistent rule use.
Siegler (1996) proposed the overlapping waves model to explain more gradual transitions between phases in development. In an overlapping waves model, each mode of behavior (or developmental phase) is represented by a wave. As waves are overlapping, the change from one mode of behavior to the next can be gradual and children may even display several modes of behavior concurrently. The extent of overlapping can vary. With development, children’s preference for a mode of behavior develops.

We contend that a restricted form of the overlapping waves model (Siegler, 1996) can describe the development of problem solving on the balance scale task. The model contains greatly overlapping waves as well as hardly overlapping waves (similar to the stairs in a staircase model). This restricted overlapping waves model is depicted in Figure 6, together with a staircase model and an overlapping waves model. All models are idealizations and describe the development of a given individual. The ages on the x-axis are derived from a multigroup LCA of data set II, with age groups featuring as group indicator. The results of this multigroup LCA clearly demonstrated the rule hierarchy that Siegler (1976, 1981) proposed. Rule I, which was mainly used by children between five and seven years old, was succeeded by Rule II, which was observed with children from eight years. Use of Rule II decreased among older children. Rule III was used by children of almost all ages, but was most frequent among children of ten years old. The onset of Rule IV was quite sudden and was first observed with children of fourteen years old. The compensation-rule was one of the most frequently used rules in children from eleven years old. The mix of the compensation-rule and Rule III was noted among children between nine and sixteen years old. Hence, the two alternative rules followed Rule III in the rule hierarchy. Finally, the SDD-rule was used by only a few children, of various ages. It should be noted that the analysis showed that children in different age groups may demonstrate similar behavior and that children of the
same age may demonstrate different rules. Hence, the values on the x-axis were chosen somewhat arbitrarily.

Figure 6 represents our hypothesis that Rule I and Rule II are non-overlapping waves and that the development from Rule I to Rule II is mainly discontinuous. Rule II and Rule III are depicted as overlapping waves. As Rule II-users improve their perception of the distance dimension, they may increasingly integrate it in their strategies. Their behavior gradually changes to the use of Rule III because they want to combine the two dimensions but do not know how. We propose that Rule III should be considered as an ensemble of strategies that each can be described as waves. Each strategy includes a combination of the distance and the weight dimension, except multiplication. The rules are used concurrently, but the preference for a certain rule changes with age. It is not clear why and how this preference changes. The transition to using Rule IV is supposed to be sudden. We contend that children shift to a higher level of formal operations (Inhelder & Piaget, 1958), or that they learn the torque-rule through instruction at school, which results in its sudden application. This rationale of the restricted overlapping waves model was based on the analyses of cross-sectional data. Of course, a longitudinal study is necessary to test the hypotheses that follow from this model.
Staircase Model

Overlapping Waves Model

Restricted Overlapping Waves Model
Figure 6. Staircase model, Overlapping waves model, and Restricted Overlapping waves model for development of reasoning on the balance scale task. The x-axis represents age in years. The y-axis represents the probability of the use of a rule. Adapted from Jansen and van der Maas (2002).

COMPUTATIONAL MODEL OF PROPORTIONAL REASONING

Several computational models have been developed to capture the developmental phenomena associated with the balance scale task. These models, which originate in different computational traditions, attempt to explain phenomena of development in proportional reasoning. So far, none of these models has been able to explain all empirical data.

Recently, van Rijn et al. (2003) proposed a computational model that is implemented in ACT-R (Anderson & Lebiere, 1998). This model can be viewed as an implementation of the restricted overlapping waves model. The validity of the model and alternative computational models was judged with regard to four empirical phenomena: 1) Stable phases and transitions: The behavior of the model should obey to Rules and the model must be able to explain the transitions between Rules. 2) Rules: a complete model should include all known empirical rules and explain the order in which they occur. 3) Transition phenomena: the sudden jump and hysteresis patterns, as demonstrated in the hysteresis experiments, should be explained. From a computational point of view it would be difficult to simulate these patterns, without using feedback about the correct response, as all existing computational models work because of feedback. 4) Torque difference effects: a computational model should explain the within item type homogeneity for moderate torque differences (Jansen & van der Maas, 1997) and within type heterogeneity for extreme torque differences (Ferretti & Butterfield, 1986). Table 7 summarizes the extend to which each existing computation model satisfies these criteria.
Table 7: Overview of empirical criteria per model.

<table>
<thead>
<tr>
<th>Empirical Criteria</th>
<th>Model Type</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Production</td>
</tr>
<tr>
<td></td>
<td>Rules</td>
</tr>
<tr>
<td>KS’78</td>
<td>√</td>
</tr>
<tr>
<td>SL’87</td>
<td>-</td>
</tr>
<tr>
<td>SL’96</td>
<td>+</td>
</tr>
<tr>
<td>McC’95</td>
<td>-</td>
</tr>
<tr>
<td>S’95</td>
<td>-</td>
</tr>
</tbody>
</table>

Note. √: Criterion fully satisfied, +: Criterion partly satisfied, -: Criterion not satisfied.


ACT-R

ACT-R is a hybrid cognitive "architecture". It is a theory for simulating and understanding human cognition (Anderson & Lebiere, 1998) in which the use of symbolic knowledge is mediated with quantitative parameters. It uses declarative memory with chunks that represent descriptive knowledge (“more weights left”) and procedural memory containing production rules in the form of IF-THEN rules (“more weights left then “left side down”). The presence of goals (like “solve the balance scale problem”) constrains the activity of the memory. The activity of each chunk or production rules is determined by its past use and success, costs, and random noise.

Development is possible since ACT-R models are able to expand their knowledge both by external input (perception) and internal modifications (declarative chunks from production
rules). Van Rijn et al (2003) also apply a new ACT-R method, called production composition (Taatgen & Anderson, 2002). Production composition creates new rules by joining production rules that occur in succession. This composition mechanism also applies to learning new behavior from declarative descriptions of problem solving actions. These declarative actions can be matched and executed by "interpretative production rules". Composed production rules have generally lower costs than the parent rules together and therefore the new production rules are favored.

Set-up of model

In the construction of the model, van Rijn et al (2003) distinguish three factors. The first factor concerns how the development proceeds, the second and third factors relate the order and timing of developmental events. Figure 7 displays the role of these factors in the set-up of the model.

The first factor underlying the behavior of the model concerns mechanisms in ACT-R, explained above, and task general knowledge. This task general knowledge is implemented as declarative representations of actions associated with the general strategy "answer a balance scale problem by searching for differences between the left and right side of the scale".

The second factor relates to task-specific concepts. It specifies when task properties like weight, distance, addition and multiplication become available. Two assumptions are made about these properties: weight precedes distance and multiplication precedes addition.

The third factor is capacity. Like in most developmental theories (for instance the Neo-Piagetian theories), van Rijn et al (2003) assume that a limited capacity initially constrains the generality of the "search for difference" strategy. Initially only one difference can be detected, later in development more than one difference can be searched for.
Figure 7: The set-up of the ACT-R model. Three factors explain the development of proportional reasoning: mechanisms, task specific concepts, and capacity constraints. Adapted from van Rijn et al. (2003).

Simulation

Exact description of the simulation of the model can be found in van Rijn et al. (2003). Roughly, the following development takes place. Initially, neither the weight nor the distance property is encoded because the associated activation levels are below threshold. Therefore, the model answers by guessing. After each guess, feedback is given about the correctness of the given answer. As the proportion of correct answers based on guessing is low, the production rules representing this strategy have a low utility. Therefore, as soon as the weight property becomes available, the model will start to incorporate this concept in the decision
process, yielding Rule I behavior. As this increases the proportion of correct answers, the production rules associated with Rule I are preferred over the pre-Rule I production rules. Since the proportion of successful answers is still relatively low, distance is incorporated shortly after it becomes available to the model. However, as the available capacity is insufficient to incorporate both properties in the decision process at the same time, the model switches its attention to distance, discarding the weight information. As determining the answers based solely on weights is less successful when the weights are equal than when the weights are unequal, the shift from weights to distances only occurs when the weights are equal. When the capacity has increased sufficiently to make it possible to use a strategy that incorporates both weights and distances, the model progresses to Rule III behavior. In this phase, the weights and distances are examined regardless of whether the number of weights is equal or not. However, as knowledge to combine the weights and distances is as yet unavailable, the model can only guess the answer if the weights and distances are both unequal. Only when the concept of addition or multiplication becomes available, is the model able to progress to the compensation-rule, and finally, to Rule IV.

The model can also be tested on series of items used in the hysteresis experiment (see Section 4). No feedback is given in this simulation. The increased perceptual saliency of the distance cue allows the model to switch from Rule I to Rule II. Under certain activation conditions, hysteresis occurs. Also torque difference effects can be simulated in this way.

Evaluation of the model

This ACT-R model is based on the evaluation of success of applied knowledge, combined with a mechanism to construct new knowledge by searching for differences between the left- and right-hand sides of presented balance scale problems. This model accounts for the main empirical phenomena as well as for the recently reported empirical
phenomena, such as learning without feedback and hysteresis.

The successful reproduction of the empirical phenomena by our model was partly realized with features that were also used in previous computational models. Like the symbolic models, behavior in the ACT-R model is based on the application of production rules, which results in rule-like behavior. As in the previous models, new production rules are learned by extending the already present knowledge. However, instead of containing a few complex production rules, our model consists of a larger number of smaller production rules. Each of these production rules executes only a small part of the complete answering process. Therefore, newly constructed production rules can simply replace older production rules instead of requiring a complex mechanism to modify existing production rules. As in the neural net models of balance scale behavior, quantitative information plays an important role in the ACT-R model (i.e., utility and activation). The dynamics of these quantitative variables are important for the description of the empirical phenomena.

The combination of features from ACT-R and the symbolic and neural net type of models provides the basis for a number of achievements specific for this model: (1) The model produces the relatively abrupt transitions, which are problematic in the non-symbolic models. (2) It explains transitions without feedback and the related transition patterns, which cannot be explained from the learning methods used in the previous models. (3) The model demonstrates that, given a non-biased training set, Rule IV will not easily be learned because of the high success-rate of the Addition Rule. (4) The presented model is able to explain both phenomena related to long-term development and phenomena that are only observable during short time-spans. (5) The model makes explicit that the notion of “search for differences” combined with a gradual increase in capacity and knowledge is sufficient to explain development on the balance scale task.
DISCUSSION

This chapter summarizes and integrates a number of recent studies on development of proportional reasoning as assessed with the balance scale task. We applied latent structure models for rule assessment, RT analysis to further analyze the rules, catastrophe theory to investigate the transition from Rule I to Rule II, and finally ACT-R modeling to test our ideas about how development proceeds on this type of tasks by simulation. Due to space limitations, we provided only a brief description of the method and results. Yet, we hope that the readers have seen enough to grasp the potential of these methods and importance of the outcomes.

First, we showed that the criticism on the concept of rules, as well as the criticism on the RAM, can be overcome by the application of LCA. In a latent class model, rules are interpreted as categorical values of a latent variable. Since a priori expectations are not necessary, new rules can be detected, with the SDD-rule as a nice example. Next, the set of rules and each rule separately can be subjected to statistical testing. This takes away the arbitrariness of traditional rule assessment methods. Using LCA, we were able to falsify the claim that a specific connectionist model shows rule-like behavior, similar to the behavior of children. We also evidenced an important assumption underlying Siegler’s rules, namely that behavior is homogeneous within item types. Finally, we reached a number of conclusions about the actual rules used by children in solving balance scale problems.

Second, we further validated the rules with an analysis of RTs. To our own surprise it was possible to use these data to extend the rule theory with predictions relating to all kinds of quantitative characteristics of balance scale items. For instance, it was possible to show that the compensation-rule is in fact a so-called buggy-rule and not an addition-rule. We also found some unexpected results. The most important was that users of Rule II seem to invest time in perceiving the distance cue also when the number of weights on the left side differs
from the number of weights on the right side. This unexpected phenomenon indicated that Rule II is very similar to Rule III. The difference is that users of Rule III guess when the two cues conflict, whereas users of Rule II apparently decide to ignore the distance cue in the choice of their response.

Third, we reported our study of the Rule I to Rule II transition. We developed a cusp model for this transition, derived unique predictions of this model and tested them in two studies. We found strong evidence for the presence of the catastrophe flags bimodality and inaccessibility and some evidence for hysteresis but we failed to find evidence for divergence. In the reaction time study, we found an indication of critical slowing down, another catastrophe flag. Although these findings require further study, it seems that this is the first time real evidence for a genuine phase transition in cognitive development is shown.

Fourth, we integrated these findings into a new general model of development of reasoning on the balance scale task. This model is a restricted version of Siegler’s (1996) overlapping waves model in which some transitions are rather sudden and others are more gradual. By analyzing transition matrices between latent class models of different parts of a balance scale test and by analyzing the relation to age, we were able to specify this overlapping waves model quite precisely, as shown in Figure 6.

Finally, we formulated a computational model of balance scale learning. One off-spin of such an attempt is that one is forced to be very explicit about the mechanisms, constraints and assumptions underlying the model. Our present model consists of ACT-R mechanisms, general knowledge, assumptions about task specific properties and the role of memory capacity constraints. We were able to explain the main empirical phenomena with this model. We think that the restricted overlapping waves model and the ACT-R model at least partly answer the "why" and "how" questions posed in the introduction. The LCA and RT analyses partly answer the "what" question. Cognitive development in this and probably many other
domains, can be characterized as a progression through a series of increasingly complex and accurate task specific rules or strategies.

This view of cognitive development is clearly consistent with other approaches to the study of development. The originality of our work lies more in the use of new advanced formal methods (statistical, mathematical and computational) than in the basic theory. Yet a few additional theoretical remarks should be made.

In the dynamical (e.g. Thelen & Smith, 1994) and in the connectionist (Bates & Elman, 1993; McClelland, 1995) accounts, the concepts of rules and symbols are often explicitly rejected. We use dynamical concepts (catastrophe theory concerns non-linear dynamical systems) and connectionist ideas (the sub-symbolic level in Act-R), yet we adopt the traditional view of rules in explaining higher mental processes (see also van der Maas, 1995). We think the results of the latent class analyses and the response time analyses support our point of view.

Our approach is generally consistent with the Piagetian and neo-Piagetian theories (e.g. Demetriou, Christou, Spanoudis & Platsidou, 2001) in that it tries to explain constructive development. A number of ideas, like the importance of cognitive capacity, perceptual cues (field dependency) and processing speed of mental procedures are adopted from these approaches. In the Act-R model, for instance, one important factor is the continuous increase in cognitive capacity or memory span. Finally, we mention that it might be possible to interpret certain ACT-R mechanisms used here as implementation of Karmilof-Smith's concept of representational redescription (Taatgen, van Rijn & Zondervan, 2003).

Of course, many new questions arise and some are still unresolved. There are at least three major steps required to test our ideas. The first is to acquire individual data over time by using a microgenetic or longitudinal design. The restricted overlapping waves model is drawn from cross-sectional data and data from the hysteresis experiment. A valid test clearly
requires longitudinal data. Second, it would be wise to validate our results with other proportional reasoning tasks, like the shadow task (Siegler, 1981). Third, the present ACT-R model was not tested on the RT data. ACT-R models are known for their ability to predict RTs. The present model must be able to mimic our results to a large extent. We see this test as important validation since the model was not constructed for this type of data simulation.

REFERENCES


481–520.


