What response times tell of children’s behavior on the balance scale task

Han L.J. van der Maas* and Brenda R.J. Jansen

Faculty of Psychology, Department of Developmental Psychology, University of Amsterdam, Roetersstraat 15, Amsterdam 1018 WB, The Netherlands

Received 9 November 2001; revised 14 March 2003

Abstract

Analysis of accuracy of responses to balance scale problems gives a global idea of the cognitive processes that underlie problem-solving behavior on this task. We show that response times (RTs) provide additional detailed information about the kind and duration of these processes. We derive predictions about the RTs from Siegler’s (1981) model for the balance scale task, including the counterintuitive prediction that young adults are slower than children in solving particular balance scale problems. The predictions were tested in a study in which 191 6- to 22-year-old participants were presented with a computerized balance scale task. RTs were analyzed with regression models. In addition to qualitative differences between items, we also modeled quantitative differences between items in the regression models. Analyses supported the predictions and provided additional knowledge on the rules. Rule II was reformulated as a rule that always involves the encoding, but not always the correct application of the distance cue. RTs provided evidence for the use of a buggy-rule and not an addition-rule. Finally, a relation between rule inconsistency and increased RT was found.

Keywords: Balance scale task; Response times; Cognitive strategies; Proportional reasoning

Introduction

Young adults are slower than children when solving balance scale problems. This statement seems very unlikely because, generally, the faster basic processing speed of
young adults makes them faster than children. The statement becomes much more likely when Siegler’s (1976, 1981) description of solution rules for the balance scale task is studied.

The balance scale is undoubtedly the best-known task used to investigate proportional reasoning. To succeed on a proportional reasoning task, it is necessary to identify the relevant task dimensions and to understand the multiplicative relation between those dimensions. The rules that children use when solving balance scale problems consist of step-like procedures that become more complex with development (Siegler, 1981). As complexity of a rule influences solution time, it seems quite likely that young adults are slower than children.

This study aims at testing the sometimes counterintuitive predictions that follow from Siegler’s rule descriptions by measuring participants’ response times (RTs) when solving balance scale problems. Accuracy data largely support the rule descriptions (Jansen & van der Maas, 1997, 2002; Siegler, 1976, 1981) but the predictions concerning RTs have not been tested extensively yet. As RTs are continuous measurements, RT data may contain additional information about the content of the cognitive processes that underlie problem solving on the balance scale task. Siegler (1989) showed, in the domain of subtraction, that RTs are very useful in validating or disconfirming strategy assessments.

Next, we exploratively study the relation between rule inconsistency and variance in RT. Rule inconsistency may be related to rule switches. In the literature on cognitive development, it is often noted that switches are accompanied with increased RTs. Alibali (1999) notes that variability is an important concept in many theoretical accounts of developmental change. She notes that, across several different theoretical frameworks and content domains, researchers observe that periods of transition are marked by high variability.

The cusp model, a mathematical model of discontinuities, which was applied to conservation of liquid quantity by Van der Maas and Molenaar (1992) and to the balance scale task by Jansen and van der Maas (2001), predicts criteria, so-called catastrophe flags, that indicate the presence of a discontinuous transition. One of these flags is the phenomenon “critical slowing down.” It implies that participants who undergo a change from one level of performance to another are very sensitive to perturbations, resulting in a strong increase in time to return to the stable performance of a rule. An increased RT after a transition is a manifestation of critical slowing down. Jansen and van der Maas (2001) applied the cusp model to the transition from Rule I to Rule II on the balance scale and tentatively concluded that this transition happens discontinuously as they found evidence for some important, necessary flags. However, Jansen and van der Maas did not record RTs and were not able to test the presence of the flag “critical slowing down.” The present dataset possibly does provide this possibility.

The results of the analyses of RTs may provide a challenge for the various symbolic (Klahr & Siegler, 1978; Langley, 1987; Sage & Langley, 1983; Schmidt & Ling, 1996; Van Rijn, van Someren, & van der Maas, 2003) and connectionist (McClelland, 1989, 1995; Shultz, Mareschal, & Schmidt, 1994; Shultz & Schmidt, 1991; Shultz, Schmidt, Buckingham, & Mareschal, 1995) models developed for the balance scale task. In the past, these computational models have been tested by use of accu-
racy data (e.g., Jansen & van der Maas, 1997; Raijmakers, van Koten, & Molenaar, 1996), which most models are able to mimic. We expect that RT data make a much more sophisticated comparison of the main computational paradigms possible. Although we will not study this hypothesis here, we will provide the target data for further comparison of the various models of balance scale behavior.

Balance scale task

On the balance scale task, a participant is asked to predict the movement of a scale. On both arms of the scale, pegs are situated at equal distances from each other and from the fulcrum. Equally heavy weights can be placed on the pegs. Participants’ responses to six types of balance scale items reveal the rule they use to solve the items. The six types can be divided into simple and conflict types. On simple-balance items, both arms of the scale hold the same number of weights, equidistant from the fulcrum. On simple-weight items, the arms contain unequal numbers of weights, equidistant from the fulcrum. Simple-distance items 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. Hence, the weight dimension and the distance dimension conflict. The scale tips to the side with the larger number of weights on conflict-weight items and tips to the side with the weights placed at the greater distance on conflict-distance items. On conflict-balance items, the scale remains in balance. An example of each item type is in Table 1. The weight-distance items and the breakdown of the conflict-balance items in types A and B will be explained below.

Many experimenters have studied children’s behavior on the balance scale task (Boom, Hoijtink, & Kunnen, 2001; Chletsos, De Lisi, Turner, & McGillicuddy-De Lisi, 1989; Ferretti & Butterfield, 1986; Ferretti, Butterfield, Cahn, & Kerkmann, 1985; Halford, Andrews, Dalton, Boag, & Zielinski, 2002; Klahr & Siegler, 1978; Kliman, 1987; Marini & Case, 1994; McFadden, Dufresne, & Kobasigawa, 1987; Normandeau, Larivée, Roulin, & Longeot, 1989; Richards & Siegler, 1981; Roth, 1991; Siegler & Chen, 1998; Surber & Gzesh, 1984; Van Maanen, Been, & Sijtsma, 1989; Wilkening & Anderson, 1982). The responses to the item types have been classified successfully into a small set of increasingly complex rules (e.g., Siegler, 1976, 1981). A rule is hypothesized to consist of consecutively executed steps. Complexity increases with development as each rule consists of the steps of the preceding rule, extended with one or more extra steps. The predictions concerning accuracy of each rule are summarized in Table 1.

It is hypothesized that the time that is involved with solving a balance scale item equals the sum of the duration of the steps that are completed. Below, we derive the required steps for each item type, given rule, from the basic model proposed by Siegler (1981). These steps are represented in Table 2.

Rule I is the simplest rule as it only involves one step, which consists of comparing the numbers of weights. The duration of this step is denoted $w$. Participants who use Rule I decide that the scale will tip to the side with the largest number of weights if the numbers are unequal and that the scale will remain in balance if the numbers are
equal. They answer simple-balance, simple-weight, and conflict-weight items correctly. The use of Rule I results in incorrect responses to simple-distance items and to the remaining conflict item types. The RTs on all item types are predicted to equal $w$.

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 if the numbers differ. However, if the numbers are equal, they also compare the distances at which the weights are placed. The duration of this step is indicated with $d$. This makes Rule II more complex than Rule I. Participants who use Rule II answer all simple items and conflict-weight items correctly, but incorrectly predict that the scale will tip to the side with the larger number of weights on conflict-distance and conflict-balance items. The RTs on items with different numbers of weights (i.e., simple-weight and conflict-items) are predicted to equal $w$, whereas the RTs on items with equal numbers of weights (i.e., simple-balance and simple-distance items) are expected to equal $w + d$ for participants who use Rule II.

---

Table 1
Predicted proportion of correct responses in the present test, given item type, for each rule

<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>![Diagram]</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Simple-weight</td>
<td>![Diagram]</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Simple-distance</td>
<td>![Diagram]</td>
<td>.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Conflict-balance B</td>
<td>![Diagram]</td>
<td>.00</td>
<td>.00</td>
<td>.33</td>
<td>.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Conflict-weight</td>
<td>![Diagram]</td>
<td>1.00</td>
<td>1.00</td>
<td>.33</td>
<td>.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Conflict-distance</td>
<td>![Diagram]</td>
<td>.00</td>
<td>.00</td>
<td>.33</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Conflict-balance A</td>
<td>![Diagram]</td>
<td>.00</td>
<td>.00</td>
<td>.33</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Weight-distance</td>
<td>![Diagram]</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td></td>
</tr>
</tbody>
</table>

*a* Answers that 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.
Rule III is more complex than Rule II because it contains two additional steps. Participants who use Rule III 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 RTs of participants who use Rule III are expected to equal $w + d$ on simple items. The first additional step is performed if both the weights and the distances are unequal. This step, indicated with $a$, includes determining whether the dimensions agree (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, indicated with $g$.

The accuracy of Rule III on conflict items is hypothesized to be at chance level. The RTs on conflict items are predicted to equal $w + d + a + g$. The duration of the last step, $g$, may provide more insight into its content. A short RT would indicate pure guessing, whereas a long RT might indicate a reasoned guess.

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. This step, indicated with $p$, includes executing the torque rule on conflict items and comparing the products of weight and distance of each side of the scale. Rule IV results in the correct response to all item types. The RTs are expected to equal $w + d$ on simple items, and to equal $w + d + a + p$ on conflict items.

Siegler (1981) points out that Rule III includes a host of idiosyncratic strategies. Here, we distinguish Rule III, which implies guessing or muddling through on conflict items, from the so-called compensation-rule (Halford et al., 2002). The term is used to refer to both the addition-rule and the buggy-rule (Van Maanen et al., 1989). The addition-rule involves the addition of weight and distance on each side of the scale and comparing the sums, in case of conflict items. The buggy-rule is also used on conflict items. It involves shifting the pile with the largest number of weights, on the smallest distance, towards the end of the scale. For each shift, a weight is taken from the shifted pile, until either the distances or the weights are equal. Note that this buggy-rule is actually an additive procedure (Ferretti et al., 1985; Jansen & van

<table>
<thead>
<tr>
<th>Item type</th>
<th>Rule model</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>$E(RT_{sb})$</td>
<td>$w$</td>
<td>$w + d$</td>
<td>$w + d$</td>
<td>$w + d$</td>
<td>$w + d$</td>
<td>$w + d$</td>
</tr>
<tr>
<td>$E(RT_{sw})$</td>
<td>$w$</td>
<td>$w$</td>
<td>$w + d$</td>
<td>$w + d$</td>
<td>$w + d$</td>
<td>$w + d$</td>
</tr>
<tr>
<td>$E(RT_{sd})$</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>
<td>$w + d + a + p$</td>
</tr>
<tr>
<td>$E(RT_{cb})$</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>
<td>$w + d + a + p$</td>
</tr>
<tr>
<td>$E(RT_{cd})$</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>
<td>$w + d + a + p$</td>
</tr>
</tbody>
</table>

Note. sb, simple-balance items; sw, simple-weight items; sd, simple-distance items; cb, conflict-balance items; cw, conflict-weight items; cd, conflict-distance items; wd, weight-distance items. 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.
der Maas, 1997, 2002; Normandeau et al., 1989). Because the response patterns of the buggy- and the addition-rules are identical, we employ one name, compensation-rule, for both rules.

The RTs of participants who use the compensation-rule are predicted to equal \( w + d \) on simple items as the correct response on these items is derived by comparing weights and distances. On conflict items, the rule involves deciding whether the greater number of weights is on the same arm as the greater distance. Participants either compare the sums of weight and distance (addition-rule) or perform buggies (buggy-rule) in the fourth step. Hence, the RT of users of the compensation-rule on conflict items is predicted to equal \( w + d + a + c \). The compensation-rule results in the correct to some conflict items.

Generating a response, consisting of motor output like positioning the mouse in the correct position and clicking the mouse, comprises a constant of the RTs of all rules. Technically, it is impossible to determine the duration of this constant, separately from the duration of the step of weight comparison, because users of all rules always execute both processes simultaneously, on any item type.

The indices in Table 2 can be conceived of as parameters. The first aim of this study is to estimate the parameters by translating these into independent variables in regression equations. The dependent variables in these equations are the RTs on each item type, for each rule.

Note that the step that consists of deciding that the dimensions agree \( (a) \) is always executed in combination with “guessing” \( (g) \), “compensation” \( (c) \), or “compare products” \( (p) \). Hence parameter \( a \) cannot be estimated. However, estimation of \( a \) can be accomplished by the introduction of an extra item type, weight-distance items. Weight-distance items are items with unequal numbers of weights, placed at unequal distances. Contrary to conflict items, the larger number of weights is also placed at a larger distance. For example, a pile of two weights is placed at the second peg from the fulcrum on the left, and a pile of four weights is placed at the fourth peg on the right arm (see the example in Table 1). According to Siegler’s (1981) model, users of complex rules solve weight-distance items with the same steps they use for solving conflict items, except the fourth step. After deciding that the greater weight is on the same side as the greater distance, in the third step, weight-distance items can be answered correctly. Hence, on weight-distance items, parameter \( a \) is isolated from the more complex steps.

It is remarkable that the model predicts this long procedure for the complex rules. It seems easy to decide that the scale will tip to the side with the larger number of weights, placed at the larger distance. Indeed, participants who use Rule I and Rule II are predicted to execute the step “compare weights” only, to derive the correct response (see Table 2). Hence, the predicted RTs on weight-distance items are predicted to contain more steps for older participants, who use complex rules, than for younger participants, who use either Rule I or Rule II. The larger number of steps is perhaps associated with longer RTs on weight-distance items in spite of the faster basic processing speed of young adults. This prediction can be used as a counterintuitive additional test of Siegler’s rule model.

The procedure used here is comparable to Donders’ (1868–1869) subtraction method. In this method, it is assumed that the time required for a mental process
can be determined by subtracting RT on a simple task from the RT on a more complex task. The difference in RT equals the duration of the process that deals with the added complexity. A general comment on this procedure concerns the assumption of an entirely serial process in which the times of the separate stages simply add. Moreover, the assumption of “pure insertion,” which implies that a stage of mental processing can be added without affecting the remaining stages (Luce, 1986), can be questioned. We think that Donders’ method is suited for analyzing RTs on balance scale problems as it is very probable that the cognitive process of solving balance scale problems indeed involves several serial stages. The presented order of successive steps corresponds to the necessary order of steps in the solution process. It is necessary to know the values of both the weight and the distance dimension, and, next, to know whether these values conflict, before a different procedure, like summing weight and distance, can be performed.

In the present study, we administered a computerized balance scale test to a sample of children and college students. Participants’ rule use was assessed by considering the response patterns. The RTs that are associated with the rules were analyzed with regression models. We report regression models at the level of rules as well as at the level of individual participants. At the level of rules, the average RTs per item, per rule, were modeled with the basic model of Siegler (1981). However, several modifications of the basic model were suggested and tested. At the individual level, each participant’s RT on each item was modeled. This level allows for modeling the influence of participants’ characteristics, like age, on RT. We first introduce the predictions for the two levels of analysis in detail.

**RT-model and predictions**

**Predictions on the level of rules**

**Basic model.** Participants’ RTs were modeled with a regression model that was derived from Siegler’s (1981) model for behavior on the balance scale task. This model represents Siegler’s (1981) basic model but also includes parameters that model the “law of practice.” This law dictates that participants gain speed during a test session because they become acquainted with the task (see, e.g., Thorndike, 1913). Heathcote, Brown, and Mewhort (2000) show that an exponential function best describes the law of practice: \( i_1 e^{-i_2 I} \), where \( I \) is the position of the item in the test.

It was tested whether it was necessary to introduce rule-independent steps. In Siegler’s (1981) model, it is suggested that all rules contain the same basic process of weight comparison. This hypothesis was translated into the basic regression model in Table 2 by using only one parameter, \( w \), across rules. Also the step “distance comparison” was estimated by using one parameter, \( d \), across rules. It was tested if these assumptions were tenable by estimating \( w- \) and \( d- \) parameters for each single rule and by comparing the fit of the complex models with the fit of the simple model.

After checking whether separate \( w- \) and \( d- \) parameters were necessary, we introduced additional parameters for the effects of quantitative properties of items. The
additional parameters, which are described below, are represented in Fig. 1. Fig. 1 is based on Siegler’s (1981) representation of rules for the balance scale task. The diamond with dashed lines and the equations between curly brackets are ad hoc additions and can be ignored for the moment.

Weight comparison. In Table 2, a step is associated with one parameter only. However, perceiving an equality of weights may take more (or less) time than perceiving an inequality of weights. It was tested whether breaking down the \( w \)-parameter into an equality parameter \( (w_0^-) \) and an inequality parameter \( (w_0^+) \) improved the description of the observed RTs. The \( w \)-parameters are indexed (starting at 0), because of the introduction of other parameters below.

Item homogeneity, responding identically to items of the same type, was proven for accuracy data (Jansen & van der Maas, 1997), but is questionable for RT data. Quantitative attributes of items, for instance, the number of weights, may influence the duration of the execution of a step. Quantitative attributes are denoted with capitals throughout.

In the case of equal numbers of weights, a large number of weights on both sides of the scale may complicate weight comparison because attention needs to be paid to each weight. The duration of the comparison of equal weights was extended with \( w^-_1 W \), where \( W \) refers to the number of weights in each pile. Hence, the duration of the comparison of equal weights was expressed as \( RT = w^+_0 + w^-_1 W \).

In the case of unequal numbers of weights, a large difference between the number of weights on the left side and the number on the right side may facilitate weight comparison because the difference is more obvious. This facilitating influence of weight difference was denoted as \( w^+_1 \Delta W \), with \( \Delta W \) referring to the difference in numbers of weight. The time that is needed to determine that the numbers of weights differ was hence modeled by the equation \( RT = w^+_0 + w^+_1 \Delta W \).

Distance comparison. Comparable hypotheses were formulated for the comparison of equal and unequal distances. The \( d \)-parameter was split up into an equality parameter \( (d^-_1 + 0) \) and an inequality parameter \( (d^+_1 + 0) \). The \( d \)-parameters include a second index (starting at 1), because of additional parameters, introduced below.

Determining that the numbers of weights are placed at different distances may be easier when the difference between the distance on the left and the distance on the right side of the scale is large and the distance dimension salient. The equation that expressed the duration of comparing unequal distances was extended with this facilitating influence of a large distance difference, resulting in the equation \( RT = d^+_1 + d^+_1 \Delta D \).

The influence of the distances at which the weights are placed on the comparison of equal distances was expected to be more complicated. The comparison may be

---

1 The introduction of \( w^-_1 \) and \( w^+_1 \) allowed for estimation of the constant, associated with generating a response. Parameters \( w^-_0 \), \( w^+_1 \), \( w^+_0 \), and \( w^+_1 \) are all associated with weight comparison. If we replace \( w^-_0 \) with constant \( m \) (referring to motor response) and hence restrict parameter \( w^-_0 \) to the value zero, \( m \) attains the value of \( w^-_0 \). The estimated value for \( w^+_0 \) becomes equal to the difference between the old value of \( w^+_0 \) and the value of \( m \) (or the old estimate of \( w^+_0 \)). The estimated values for the mediating parameters \( w^-_1 \) and \( w^+_1 \) remain the same.
Fig. 1. Rules constructed from steps with all parameters assumed to be relevant. Elements drawn with dashed lines and parameters between curly brackets are not part of the predicted model. Deleted parameters did not differ significantly from zero. \( m, \) motor response; \( y_1, \) age-parameter; \( i_1 \) and \( i_2 \) are the exponent and the base, respectively, for the law of practice; \( I, \) position of item in the test; \( \text{incr}, \) rule inconsistency. \( w_0 \) and \( w_1 \) concern the comparison of equal weights; \( w_0 \) and \( w_1 \) concern the comparison of unequal weights. \( W, \) number of weights on each side of the scale; \( \Delta W, \) difference between the number of weights on the left and the number of weights on the right side of the scale. \( d_{20}^c \) and \( d_{21}^c \) concern the comparison of equal distances, whereas \( d_{20} \) and \( d_{21} \) concern the comparison of unequal distances for Rule II; \( d_{20}^c \) and \( d_{21}^c \) concern the comparison of equal distances, whereas \( d_{20}^c \) and \( d_{21}^c \) concern the comparison of unequal distances for Rules III, the compensation-rule and Rule IV. \( D, \) distance on each side of the scale; \( \Delta D, \) difference between the distance on the left and the distance on the right side of the scale; \( a, \) deciding whether the weight and distance dimension agree. Parameters \( g_0 \) and \( g_1 \) are related to Rule III. Parameter \( c_0 \) concerns the compensation-rule. Parameters \( b_1 \) and \( b_2 \) concern the buggy-rule. \( B, \) number of buggies; \( \text{max}(W), \) number of weights in the largest pile. Parameters \( s_1, s_2, \) and \( s_3 \) concern the addition-rule. \( S, \) total of the sums of weight and distance on each side of the scale; \( \text{Sym}, \) presence of symmetries in an item; \( \Delta S, \) difference between the sums. Parameters \( p_0, p_1, p_2, p_3, \) and \( p_4 \) relate to Rule IV. \( \sum P, \) sum of products of weight and distance on each side of the scale ("product difficulty"); \( \Delta P, \) product difference; \( H, \) hesitation. See text for explanation of the exact meaning of the parameters.
easy when the piles of weights are both placed at the far ends of the scale or close to the fulcrum. In these cases, counting is redundant to determine that the distances of the piles are equal. The comparison may be more complicated when the weights are placed near the middle of each arm of the scale. For ease of notation, the effects of the distance of the piles from the middle of the arm was denoted as \( d \).\(^2\) The duration of the comparison of equal distances was expressed in the equation \( RT = d_{0} + d_{1}D \).

The additional equations for the process of distance comparison are shown in Fig. 1. The indices of the parameters associated with distance comparison are different for Rule II because of the introduction of additional parameters below.

**Compensation.** Testing the effects of quantitative attributes allows for the study of the content of the compensation-rule because the hypothesized effects differ for the addition- and the buggy-rule. For the buggy-rule, we hypothesized that the number of buggies \((B)\) was related to \( RT \). The number of buggies was defined as the minimum of the weight and the distance difference of an item \((B = \min(\Delta W,\Delta D))\). The influence of the number of buggies was denoted as \( b_1B \). Also, we expected that performing buggies was more difficult if the number of weights in the pile that needed to be shifted (i.e., the pile with the largest number of weights) was large. Shifting the pile with a large number of weights may be more difficult because this demands more (memory) effort than shifting a small number of weights. This effect was noted as \( b_2 \max(W) \).

For users of the addition-rule, larger sum sizes may be associated with longer \( RT \).s. The lengthening effect of large sums was denoted as \( s_1S \), with \( S \) representing the total sum of all values on the weight and the distance dimension. A large difference between the sum of weight and distance on the left side and the sum of weight and distance on the right side of the scale \((\Delta S)\) may be associated with shorter \( RT \)s for users of the addition-rule because large sum differences perhaps facilitate items. The effect of sum difference was denoted as \( s_2\Delta S \).

We employ the term “symmetry” \((\text{Sym})\) on conflict items to refer to equal values of the weight dimension on one side and of the distance dimension on the other side. In the conflict-balance A item in Table 1, the number of weights on the left side of the scale equalizes the distance on the right side and vice versa. The conflict-distance item in Table 1 also includes a symmetry: the distance on the left side is equal to the number of weights on the right side. Recognizing a symmetry perhaps simplifies an item. The facilitating influence of symmetries in a conflict item on \( RT \) was expressed as \( s_3\text{Sym} \). Summarized, the equation expressing the (extra) time that is needed for solving a conflict item with the compensation-rule was \( RT = c_0 + b_1B + b_2 \max(W) + s_1S + s_2\Delta S + s_3\text{Sym} \), where \( c_0 \) represents the intercept of the equation.

\(^2\) The actual formula for \( D \) is: \( D = \left|2.5 - D_{1}\right|\) where \( D_{1}\) refers to the distance from the fulcrum at which the weights are placed. In case of equal distances, the distance of the pile on the left equals the distance on the right side. The distance between the middle of the arm and the fulcrum is 2.5 pegs as the scale we used had four pegs on each arm. Hence, \( \left|2.5 - D_{1}\right| \) represents the distance from the middle of an arm. The maximum distance from the middle is obtained when a pile of weights is placed on the farthest \((D_{1} = 4)\) or the closest peg from the fulcrum \((D_{1} = 1)\) and equals 1.5.
**Product comparison.** The (extra) time involved with solving a conflict item with Rule IV was expressed as $RT = p_0 + p_1 Sym + p_2 \sum P + p_3 \Delta P$. The first parameter, $p_0$, refers to the intercept of the equation. Parameter $p_1$ is associated with the mediating effect of the presence of symmetries in a conflict item. The slowing influence of product difficulty, the size of the products that need to be calculated, is expressed in $p_2$. Parameter $p_3$ expresses the possible facilitating effect of a large difference between the products (Ferretti et al., 1985).

**Guessing.** Possible mediators of the RT on conflict items for users of Rule III are weight difference, distance difference, the saliency of each dimension, and the sum of weight difference and distance difference. Only the parameters associated with the sum of weight difference and distance difference $(g_0 + g_1(\Delta W + \Delta D))$ are shown in Fig. 1.

**Predictions on the individual level**

**Age.** The individual level allowed for testing participant characteristics that may explain differences between RTs of participants using the same rule. The first participant characteristic, next to rule, was “Age.” The term that is associated with the effect of age is $y_1 Y$, where $Y$ corresponds to the years of age.

**Age interaction.** Age may also influence the separate processes of weight comparison, distance comparison, guessing, product comparison, and compensation. These process-specific age-interactions are not indicated in Fig. 1 but were tested.

**Rule inconsistency.** The second additional participant characteristic was the consistency with which a participant performed a rule. Although many findings demonstrate consistent use of rules, rule inconsistency is also observed (Jansen & van der Maas, 2002; Siegler, 1989). Here, rule inconsistency refers to the deviation between the response pattern that was associated with the rule that the participant was expected to use and his observed response pattern. Rule inconsistency may be related to rule switches.

Jansen and van der Maas (2002) included the hypothesis on the discontinuous transition from Rule I to Rule II in a model for development of reasoning on the balance scale task. The model predicts the alternated use of Rule II, Rule III, and the compensation-rule and therefore predicts gradual transitions between these rules. The transition to Rule IV is expected to be sudden. Once participants have learned this rule, which always results in the correct response, they do not switch to other rules. Based on this developmental model, we expected that rule inconsistency was associated with increased RTs, especially for participants who use Rule I and Rule II.

**Method**

**Participants**

A total of 147 children of two primary schools (from the northwest and from the center of the Netherlands) and of one secondary school (from the center of the Neth-
erlands) participated in the study. Schools were recruited by telephone and sent a letter with information about the experiment. Parents were asked for written permission for their child’s participation. The final sample was comprised of participants ranging from age 6 to 15 years, with sample size, by age, respectively of 13, 26, 22, 15, 21, 18, 12, 12, 6, and 1. The sample contained 1 child of unknown age. Included were 74 boys and 73 girls. Also a total of 44 undergraduate psychology students of the University of Amsterdam participated in the experiment to meet course requirements. Included were 9 male and 35 female participants. The sample was comprised of college students ranging from age 18 to 25, with sample size, by age, respectively of 5, 12, 9, 7, 6, 1, 1, and 1. The age of two college students was unknown.

The data were gathered in two consecutive years. The children in (Dutch) grades 3, 4, and 5 (average age was 8.06 years, $SD = .93$) and the college students were tested in the first year, whereas the children in grades 6, 7, and 8 (average age was 11.17 years, $SD = .85$) were tested in the second year. The tests were identical, but weight-distance items were added in the second year.

**Material**

The administered computer test included the six types of balance scale items that Siegler (1976, 1981) designed and the weight-distance items. The conflict-balance items were divided into subtypes A and B. The response of the compensation-rule is correct on conflict-balance A items and incorrect on conflict-balance B items. Furthermore, the construction of the items in this test was such that the compensation-rule resulted in an incorrect response to all conflict-weight items and in the correct response to all conflict-distance items. The test consisted of 10 sets of the seven item types. The types were arranged in the same order in each set: simple-balance, simple-weight, simple-distance, conflict-balance B, conflict-weight, conflict-distance, and conflict-balance A. The order was held constant because random order (but without items of the same type in succession) is difficult to obtain and fixed order leaves the possibility for testing for some order effects. The first five blocks contained unique items. The last five blocks contained the same items as the first five blocks, but mirrored such that the configuration of weights on the left side was transferred to the right side of the scale and vice versa. However, different simple-balance items were constructed for the last five blocks because mirroring an simple-balance item results in duplication of the item. Four practice items preceded the test. On these items, weights were placed at only one side of the scale. Six weight-distance items followed the test in the second year of the study.3 Table 1 summarizes the predicted proportion of correctly answered items, given each item type, for each rule.

The test was presented on a Macintosh Power PC, 4400. The purple colored balance scale, as presented on the computer monitor, was 5.11 cm high and 18 cm wide. Each arm of the scale had four pegs, separated 1.94 cm from each other and the fulcrum. The height of each peg was 2.47 cm. The scale could tilt to the right, tilt to the

---

3 Details of the items can be attained from the first author.
left, and remain in balance. Two brown blocks, 2.28 cm high and 2.43 cm wide, underneath the arms of the scale prevented the scale from tipping. Black weights, .25 cm × 1.52 cm, were placed at the pegs. An empty space of .21 cm between stacked weights facilitated counting of the weights. The maximum number of weights on a peg was six.

**Procedure**

The test was designed to allow participants to work alone, but an assistant aided each participant during the introduction and the practice phase and monitored the progress during the test phase. The children were tested individually in a separate classroom, within their school, whereas the college students were tested in a quiet room at the university. Although three participants were tested in the same room, they were not allowed to communicate with each other.

For the children in grades 3, 4, and 5, the introduction, described below, consisted of an animation of a female doll. All questions and remarks of the explanation (in Dutch) were printed on the screen, which the assistant read aloud. A new display appeared after the participant clicked on a yellow arrow, in the lower right corner of the screen, with the mouse. The introduction was similar, but did not contain the doll, for the children in grades 6, 7, and 8. For the college students, the introduction also did not contain the doll and the printed text was not read aloud by the assistant.

In the first sequence of displays, the doll asked for the participant’s name and explained that the balance scale was a seesaw. She made it tilt by jumping alternately on the left and the right side of the scale. Horizontal, red arrows of the same width appeared between the pegs to illustrate that all pegs were situated at equal distances from each other. Next, the doll showed the two brown blocks, which were placed underneath the scale. She demonstrated that the scale could no longer seesaw, by jumping on the left side of the scale, while the scale remained in balance. The doll showed that weights could be put on the pegs and that all weights were of equal weight.

The next sequence of displays showed that weights were placed on the scale according to the configuration of the first practice item. The participant was asked to predict what would happen if the supporting blocks were removed. In the next display, the three response possibilities appeared: a button with a scale tilting to the left, a button with a scale in equilibrium, and a button with a scale tilting to the right. The participant was asked to click on the button of her choice with the mouse. This button lighted up and the other two buttons disappeared. Fig. 2 contains a display of a conflict-distance item and the three response buttons.

The participant was asked to respond to all practice items. The assistant checked whether the participant understood that she had to imagine what would happen if the blocks, that supported the scale, were removed. Some children tended to answer that the scale remained in balance on all items, maybe because they thought that the supporting blocks prevented the scale from tipping. The test phase was announced after the participant responded correctly to three of the four practice
items. The assistant did not help the participant during the test phase. The test took 18 min on average. The participants did not get any instructions on speed of responding.

Data analysis

Rule use. Participants’ response patterns were classified into rules with a rule assessment methodology comparable to Siegler’s (1976, 1981) and with cluster analysis. Weight-distance items were excluded from rule classification because participants’ responses to these items were studied exploratively only and because only a subset of the participants responded to these items. In the rule assessment methodology, participants’ response patterns were compared to the expected response patterns of Rule I, Rule II, the compensation-rule, and Rule IV. The proportion of corresponding responses between theoretical and observed response patterns was considered for each participant, for each rule. If the maximum proportion of corresponding responses exceeded the criterion of .75, the participants’ response pattern was classified to the rule concerned. The criterion of .75 seems more tolerant than the criterion of .83 (20 out of 24 items according to a rule) that is used in Siegler’s (1981) original work. However, a criterion of .75 is sufficient in view of the much larger number of items (70).

A participant’s response pattern was classified into Rule III if the proportion of correctly answered simple items was higher than .75 and the proportion of correctly answered conflict items was lower than .75 because the use of Rule III results in a correct response to simple items and to guessing on conflict items. A response pattern was classified into a class that was identified with “unknown” if both the proportion of correctly answered simple items and the proportion of correctly answered conflict items were lower than .75.

An iterative cluster analysis over participants was also performed to classify participants into rules. Iterative cluster analysis can be used to cluster participants with similar scores together. The resemblance between participants of the same cluster is
high, whereas the resemblance between participants of different clusters is low. Cluster analysis was performed on the numbers of correct simple-balance, simple-weight, simple-distance, conflict-balance B, conflict-weight, conflict-distance, and conflict-balance A items. In previous research, we applied latent class analysis to assess rules (Jansen & van der Maas, 1997, 2001, 2002). However, the application of latent class analysis is complicated with this large set of items. Cluster analysis is utilized because it is related to the technique of latent class analysis (McCutcheon, 1987) and can be applied to sum scores.

*RTs and rules.* The observed RTs were modeled by means of regression models. Whether the addition of parameters resulted in a significant improvement of fit of the model, compared to the loss of degrees of freedom, was tested by means of an ANOVA-test. All analyses were performed with R (Bates & Watts, 1988). The parameters of the best-fitting model were used to predict the RTs on the weight-distance items.

**Results**

*Descriptive statistics*

Table 3 displays the average RT and the average proportion of correctly answered items, by item type, for five age groups. Age groups were defined such that the groups contained about equal numbers of participants. The average RTs on simple-balance and simple-weight items were quite stable over age groups, but low for college students. The RT on simple-distance items was longer for the older children. The RTs on conflict items show a large difference between the short RTs of the children in the youngest age groups (from 6 to 9 years old) and the long RTs of the older children (from 10 to 15 years old) and the college students.

The average proportions of correctly answered simple-balance and simple-weight items were high for all age groups. This proportion clearly increased over age for simple-distance items. Also the proportion of correctly answered conflict items increased, except for one item type: the proportion clearly decreased for conflict-weight items.

*Rule use*

By means of the rule assessment methodology, 60 participants were classified into Rule I, 43 participants were classified into Rule II, and 34 participants were classified into Rule III. The response patterns of 15 participants matched the pattern of Rule IV, whereas the patterns of 26 participants matched the pattern of the compensation-rule. The response patterns of 12 participants did not match any of the theoretical patterns and were hence classified as “unknown.” For one participant, the proportion of matching responses with Rule I was equal to the proportion of matching responses with Rule II (both were .93). This response pattern was classified in Rule I, because the frequency of this rule was higher.
Iterative cluster analysis was performed to verify the rule assessment methodology solution. The average variance within clusters was used to decide on the optimal number of clusters. Increasing the number of clusters from 2 to 3, 4, 5, and 6 resulted in a decrease of the average within cluster variance of 8.72, 6.19, 4.59, and 2.49, respectively. The decrease was only 1.59, 1.38, 1.17, and .68 for the extension to 7, 8, 9, and 10 clusters, respectively. Hence, the within cluster variance decreased substantially when the number of clusters was increased until it equaled six, but decreased less after more clusters were added. Moreover, the addition of clusters would probably result in particular clusters containing small numbers of participants. Table 4 shows the proportion of correct responses on each of the seven item types for each of the six clusters and includes information on the ages of the participants in the various clusters.

The clusters could easily be interpreted in terms of rules (compare Table 2). Only the sixth cluster, consisting of 10 children, was not identified and was excluded from the RT analyses.

The categorization based on cluster analysis contained all important known rules (e.g., Jansen & van der Maas, 1997, 2002; Siegler, 1976) and was very similar, in

### Table 3
Average response time and average proportion of correct responses by age group

<table>
<thead>
<tr>
<th>Item type</th>
<th>Age in years</th>
<th>6–7 (n = 39)</th>
<th>8–9 (n = 37)</th>
<th>10–11 (n = 39)</th>
<th>12–15 (n = 31)</th>
<th>18–25 (n = 44)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average RT in s (SD in parentheses)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sb</td>
<td></td>
<td>3.46 (1.04)</td>
<td>2.95 (.67)</td>
<td>3.09 (.76)</td>
<td>3.13 (1.17)</td>
<td>2.35 (.44)</td>
</tr>
<tr>
<td>sw</td>
<td></td>
<td>3.48 (1.69)</td>
<td>2.72 (.75)</td>
<td>3.22 (1.12)</td>
<td>2.99 (.99)</td>
<td>2.34 (.93)</td>
</tr>
<tr>
<td>sd</td>
<td></td>
<td>3.92 (1.45)</td>
<td>3.59 (1.27)</td>
<td>4.25 (1.61)</td>
<td>4.30 (1.25)</td>
<td>3.04 (1.39)</td>
</tr>
<tr>
<td>cb-B</td>
<td></td>
<td>3.92 (2.16)</td>
<td>3.65 (1.78)</td>
<td>5.09 (2.91)</td>
<td>6.29 (2.77)</td>
<td>5.23 (1.78)</td>
</tr>
<tr>
<td>cw</td>
<td></td>
<td>3.86 (1.46)</td>
<td>3.47 (2.02)</td>
<td>4.5 (2.12)</td>
<td>6.66 (3.91)</td>
<td>4.87 (1.73)</td>
</tr>
<tr>
<td>cd</td>
<td></td>
<td>4.11 (2.32)</td>
<td>3.83 (1.57)</td>
<td>5.11 (2.53)</td>
<td>5.90 (2.34)</td>
<td>4.30 (1.77)</td>
</tr>
<tr>
<td>cb-A</td>
<td></td>
<td>3.62 (1.61)</td>
<td>3.44 (1.71)</td>
<td>4.72 (2.55)</td>
<td>6.73 (2.84)</td>
<td>4.64 (1.67)</td>
</tr>
<tr>
<td>wd</td>
<td></td>
<td>—a</td>
<td>3.08 (1.82)</td>
<td>2.97 (1.24)</td>
<td>3.94 (2.41)</td>
<td>—a</td>
</tr>
<tr>
<td>Average proportion correct (SD in parentheses)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sb</td>
<td></td>
<td>.96 (.09)</td>
<td>.98 (.06)</td>
<td>.99 (.04)</td>
<td>1.00 (.02)</td>
<td>.98 (.06)</td>
</tr>
<tr>
<td>sw</td>
<td></td>
<td>.93 (.22)</td>
<td>.95 (.17)</td>
<td>.94 (.18)</td>
<td>.97 (.13)</td>
<td>.97 (.11)</td>
</tr>
<tr>
<td>sd</td>
<td></td>
<td>.16 (.31)</td>
<td>.47 (.46)</td>
<td>.57 (.40)</td>
<td>.80 (.32)</td>
<td>.93 (.12)</td>
</tr>
<tr>
<td>cb-B</td>
<td></td>
<td>.02 (.08)</td>
<td>.04 (.17)</td>
<td>.11 (.20)</td>
<td>.26 (.32)</td>
<td>.41 (.33)</td>
</tr>
<tr>
<td>cw</td>
<td></td>
<td>.90 (.23)</td>
<td>.91 (.23)</td>
<td>.78 (.32)</td>
<td>.51 (.33)</td>
<td>.51 (.35)</td>
</tr>
<tr>
<td>cd</td>
<td></td>
<td>.11 (.26)</td>
<td>.15 (.28)</td>
<td>.32 (.34)</td>
<td>.64 (.33)</td>
<td>.82 (.28)</td>
</tr>
<tr>
<td>cb-A</td>
<td></td>
<td>.02 (.05)</td>
<td>.05 (.20)</td>
<td>.17 (.28)</td>
<td>.34 (.30)</td>
<td>.53 (.35)</td>
</tr>
<tr>
<td>wd</td>
<td></td>
<td>—a</td>
<td>—a</td>
<td>1.00 (.00)</td>
<td>.95 (.13)</td>
<td>—a</td>
</tr>
</tbody>
</table>

Note. sb, Simple-balance items; sw, simple-weight items; sd, simple-distance items; cb-B, conflict-balance B items; cw, conflict-weight items; cd, conflict-distance items; cb-A, conflict-balance A items; wd, weight-distance items.

a Weight-distance items were only administered in the second year of the study.
b n = 4.
c n = 37.
number and interpretation of the categories, to the categorization based on rule assessment methodology. Participants who mainly answered “balance” on all balance scale items were in the unidentified group, according to both the rule assessment methodology classification and the cluster classification. The resemblance between the two categorizations was high, considering the high value of .85 for Cohen’s Kappa and the high percentage, 88%, of identical classifications. The cluster analysis categorization was preferred because it was based on similarities and differences in the data and not on pre-specified rules.

The hierarchy observed by Siegler (1981) was also observed here. The compensation-rule is probably a variant of Rule III as the average ages of the participants using either rule were close to another. Note that the standard deviations were almost four years, except for Rule I. Hence, children of the same age may use different rules, and users of the same rule may differ remarkably in age.

**Overview of the data**

The continuous lines in Fig. 3 represent the observed average RTs by item type by rule. Vertical lines indicate standard errors. Dashed and dotted lines represent models that are explained below. The observed RTs roughly demonstrated the expected patterns, based on Siegler’s (1981) basic model (see Table 2). Rule I-users responded equally fast to all item types. Rule II-users were expected to respond fast to simple-weight and conflict items and slowly to simple-balance and simple-distance items. Fig. 3 shows that Rule II-users were indeed faster on simple-weight items than on simple-balance items. Contrary to the expectation that the RTs on conflict items were equal to the RT on simple-weight items and shorter than the RT on simple-balance items, the observed RTs on conflict items were much longer than the RTs on both types of items. The average RT on conflict-distance items was even longer than the average RTs on the other conflict items.
Participants who used Rule III, the compensation-rule, or Rule IV were expected to show short RTs on simple and long RTs on conflict items. This pattern was indeed observed for Rule III-users. However, there was also a difference within the simple item types: the average RT on simple-distance items was longer than the average RTs on simple-balance and simple-weight items. Rule IV-users also showed the expected pattern. However, their RTs on conflict-distance and conflict-balance A items were considerably shorter than their RTs on conflict-balance B and conflict-weight items. Participants who used the compensation-rule also showed short RTs on simple and long RTs on conflict item types. Again, the average RT on simple-distance items was longer than the average RTs on simple-balance and simple-weight items. Next, the average RT on conflict-distance items was shorter than the average RT on other conflict items.

**Analyses of RTs on the level of rules**

All RTs, including those that were associated with responses that were inconsistent with the rule used by the participant, were analyzed. Inspection of the RT dis-
tributions showed that there were many extreme outliers. Trimming with 7.5% on both the “slow” and the “fast” RTs was chosen to exclude most of these outliers. Trimming with an even higher percentage made the computation of means and standard deviations (especially of the weight-distance items) unreliable.

The analyses at the level of rules included 5 (rules) × 70 (items) = 350 data points. Continuous lines in Fig. 4 represent the observed data, organized by item type and by rule. Dotted and dashed lines refer to models that are described below. Fig. 4 shows both between item type variation (like in Fig. 3), and within item type variation. The law of practice was visible for all item types and all rules. However, the law of practice cannot explain the long RTs of some items that were presented late in the test. For instance, the average RT of Rule III-users on the eighth conflict-distance item was quite long. Within item type variation was most apparent on conflict items, for users of complex rules. Some patterns of within item type variation were noted across rules: the RT of all rules on the fourth simple-balance item was rather short, but rather long on the ninth simple-balance item. The pattern of RTs on conflict-distance items was quite similar for Rule III and the compensation-rule.

**Basic model.** Model 1 contained the parameters of the basic model in Table 2 and of the equation that models the law of practice. The explained variance of model 1 was .72. The estimates of model 1 are in the second column of Table 5. The duration of the step of comparing weights \((w)\) was estimated at 2.75 s, \(t = 27.78, p < .01\). Parameter \(d\) (distance comparison) was estimated at \(-.22\) s, \(t = -2.24, p < .05\). A negative value is inconsistent with the subtraction method: an extra process is expected to increase RT, not to lower it. Perhaps \(d\) was estimated at a negative value because of the unexpected pattern of RTs of Rule II. Parameter \(c\) (“compensation”) was estimated at 2.39 s, \(t = 17.76, p < .01\), parameter \(p\) (“cross products equal?”) was estimated at 2.62 s, \(t = 19.43, p < .01\), and parameter \(g\) (“guess”) was estimated at 2.28 s, \(t = 16.89, p < .01\). The law of practice was modeled as \(1.83e^{-0.07t}\).

Dotted lines in Fig. 3 represent the predicted values for model 1. The negative estimate of parameter \(d\) obscured the predicted values in two ways. First, it was hypothesized that the average RT on simple items was higher for users of Rule III, the compensation-rule, and Rule IV than for users of Rule I and Rule II because the complex rules included two steps on these items, whereas the simple rules included only one step. Contrarily, Fig. 3 shows that the predicted values are longer for the simple rules than for the complex rules. Second, the average RTs on simple-balance and simple-distance items were hypothesized to be longer than the average RTs on the remaining item types, for Rule II-users (see Table 2). However, the predicted values of the RTs on simple-balance and simple-distance items were shorter than the RTs on the remaining item types due to the negative value of parameter \(d\).

The predicted values for model 1 were also represented in Fig. 4, again with dotted lines. Although the predicted values roughly followed the observed RTs, Fig. 4 clearly shows that a lot of variation remained unexplained. For instance, the observed RTs on conflict items for the complex rules clearly showed much more variance between rules and item types than the predicted RTs.
Fig. 4. RTs, given item, given rule. sb, simple-balance; sw, simple-weight; sd, simple-distance; cbB, conflict-balance B; cw, conflict-weight; cd, conflict-distance; cb-A, conflict-balance A; RIIIc, compensation-rule. The x-axis represents the set or block of items in the test. The y-axis represents RT, measured in seconds. Continuous lines depict observed average RTs. Dotted lines represent model 1; dashed lines represent model 2a. Details of the models are in the text and in Table 5. The average standard error of the RTs was .30.
Rule-independent steps. In model 1a, a general \( w \)-parameter and separate additional \( w \)-parameters for Rules II, III, IV, and the compensation-rule were estimated. Only the estimated value for the additional \( w \)-parameter for Rule II, .39, \( t = 2.94, p < .01 \), was significant. In model 1b, the same test was performed for the step “distance comparison.” One general \( d \)-parameter and additional \( d \)-parameters for Rule III, Rule IV, and the compensation-rule were estimated. No \( d \)-parameter was estimated for Rule I as this rule does not involve distance comparison. It turned out that the general distance comparison parameter was estimated at a value that was significantly too low for Rule III, Rule IV, and the compensation-rule. Again, Rule II was the cause of this effect. We introduce separate parameters for Rule II below (Model 3). With the exception of Rule II, the assumption of rule-independent \( d \)- and \( w \)-parameters can be accepted.

Model 2. Model 2 was based on model 1 but consisted of the parameters that were introduced in the Predictions section and represented in Fig. 1. The diamond with dashed lines and the parameters between curly brackets were not part of model 2 and are explained later. The explained variance of model 2 was .85. An ANOVA test revealed that the extra parameters of model 2 significantly improved the description of the observed data, compared to model 1, \( F(15, 328) = 18.007, p < .01 \). Table 5 contains the estimated values of the parameters of model 2. Some parameters were estimated at non-significant values. These parameters, except the so-called intercepts, were deleted and only parameters that were estimated at significant values were accepted in model 2a. The predicted values of model 2a are given in Fig. 4, with dashed lines. Both the explained variance (.85) and the fit of model 2a were comparable to that of model 2, \( F(4, 328) = 1.222, p = .30 \). Each set of parameters, as introduced in the Predictions section, is clarified separately below.

Weight comparison. The comparison of equal weights was modeled with parameters \( w = 0 \) and \( w = 1 \). Parameter \( w = 0 \) served as intercept, whereas parameter \( w = 1 \) indicated the increase in RT when the number of weights was increased. The basic time for comparing numbers of weights, \( w = 0 \), was 2.00, \( t = 11.33, p < .01 \). Adding one weight in each pile increased inspection time by .18 s, \( t = 2.07, p < .01 \). Comparing unequal weights was modeled with parameters \( w \neq 0 \) and \( w \neq 1 \). The duration of comparing unequal numbers of weights was estimated at 2.97 s, \( t = 22.98, p < .01 \). Increasing the difference between the number of weights by one decreased RT .12 s, \( t = -2.17, p < .05 \).

Distance comparison. Comparing equal distances was modeled with parameters \( d^{=}_{1.0} \) and \( d^{=}_{1.1} \). Parameter \( d^{=}_{1.0} \) was estimated at the non-significant value of .34, \( t = 1.42, p = .16 \), whereas parameter \( d^{=}_{1.1} \) was estimated at \( -.55, t = -3.35, p < .01 \), indicating that RT was shorter if the piles of weight were close to the ends of the scale or near the fulcrum. Comparing unequal distances was modeled with parameters \( d^{=}_{1.0} \) and \( d^{=}_{1.1} \). Parameter \( d^{=}_{1.0} \) was estimated at .84, \( t = 2.42, p < .01 \), whereas parameter \( d^{=}_{1.1} \) was estimated at \( -.35, t = -3.82, p < .01 \). Hence, the time that was spent on the comparison of unequal distances decreased if distance difference increased.

Deciding whether the dimensions conflict or agree. Parameter \( a \) was estimated at \( .86, t = 2.50, p < .05 \), and corresponded to the time that was needed to decide whether the values on the weight dimension and the values on the distance dimension
Table 5

Estimated values of the parameters of regression models at the rule level

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Model</th>
<th>1&lt;sup&gt;a&lt;/sup&gt;</th>
<th>2&lt;sup&gt;a&lt;/sup&gt;</th>
<th>3&lt;sup&gt;c&lt;/sup&gt;</th>
<th>4&lt;sup&gt;b&lt;/sup&gt;</th>
<th>5&lt;sup&gt;b&lt;/sup&gt;</th>
<th>5&lt;sup&gt;b&lt;/sup&gt;, age interactions</th>
</tr>
</thead>
<tbody>
<tr>
<td>w&lt;sub&gt;0&lt;/sub&gt;</td>
<td>2.75&lt;sup&gt;c&lt;/sup&gt;</td>
<td>2.00</td>
<td>2.02</td>
<td>3.18</td>
<td>3.39</td>
<td></td>
<td>Comparison of equal weights</td>
</tr>
<tr>
<td>w&lt;sub&gt;1&lt;/sub&gt;</td>
<td>.18</td>
<td>.18</td>
<td>.22</td>
<td>.18</td>
<td>.004&lt;sup&gt;b&lt;/sup&gt;</td>
<td></td>
<td>Comparison of equal weights</td>
</tr>
<tr>
<td>w&lt;sub&gt;2&lt;/sub&gt;</td>
<td>2.97</td>
<td>2.60</td>
<td>4.08</td>
<td>4.28</td>
<td></td>
<td></td>
<td>Comparison of unequal weights</td>
</tr>
<tr>
<td>w&lt;sub&gt;3&lt;/sub&gt;</td>
<td>-.12</td>
<td>-.03&lt;sup&gt;b&lt;/sup&gt;</td>
<td>-.05&lt;sup&gt;b&lt;/sup&gt;</td>
<td>-.07&lt;sup&gt;b&lt;/sup&gt;</td>
<td>.001&lt;sup&gt;b&lt;/sup&gt;</td>
<td></td>
<td>Comparison of unequal weights</td>
</tr>
<tr>
<td>d&lt;sub&gt;1,0&lt;/sub&gt;</td>
<td>-.22&lt;sup&gt;c&lt;/sup&gt;</td>
<td>.34&lt;sup&gt;b&lt;/sup&gt;</td>
<td>.15&lt;sup&gt;b&lt;/sup&gt;</td>
<td>1.29</td>
<td>1.28</td>
<td></td>
<td>Comparison of equal distances, Rule III, compensation-rule, IV</td>
</tr>
<tr>
<td>d&lt;sub&gt;1,1&lt;/sub&gt;</td>
<td>-.55</td>
<td>-.51</td>
<td>-.62</td>
<td>-1.08</td>
<td>.03&lt;sup&gt;b&lt;/sup&gt;</td>
<td></td>
<td>Comparison of equal distances, Rule III, compensation-rule, IV</td>
</tr>
<tr>
<td>d&lt;sub&gt;1,0&lt;/sub&gt;</td>
<td>.84</td>
<td>.77</td>
<td>1.95</td>
<td>1.95</td>
<td></td>
<td></td>
<td>Comparison of unequal distances, Rule III, compensation-rule, IV</td>
</tr>
<tr>
<td>d&lt;sub&gt;1,1&lt;/sub&gt;</td>
<td>-.35</td>
<td>-.41</td>
<td>-.49</td>
<td>-.67</td>
<td>.01&lt;sup&gt;b&lt;/sup&gt;</td>
<td></td>
<td>Comparison of unequal distances, Rule III, compensation-rule, IV</td>
</tr>
<tr>
<td>d&lt;sub&gt;0,0&lt;/sub&gt;</td>
<td>.70&lt;sup&gt;b&lt;/sup&gt;</td>
<td>1.00</td>
<td></td>
<td>1.01</td>
<td></td>
<td></td>
<td>Comparison of equal distances, Rule II</td>
</tr>
<tr>
<td>d&lt;sub&gt;1,0&lt;/sub&gt;</td>
<td>-.66</td>
<td>-.67</td>
<td>-1.10</td>
<td>.04&lt;sup&gt;b&lt;/sup&gt;</td>
<td></td>
<td></td>
<td>Comparison of equal distances, Rule II</td>
</tr>
<tr>
<td>d&lt;sub&gt;0,1&lt;/sub&gt;</td>
<td>.04&lt;sup&gt;b&lt;/sup&gt;</td>
<td>.19&lt;sup&gt;b&lt;/sup&gt;</td>
<td>.20&lt;sup&gt;b&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
<td>Comparison of unequal distances, Rule II</td>
</tr>
<tr>
<td>d&lt;sub&gt;1,1&lt;/sub&gt;</td>
<td>.34</td>
<td>.34</td>
<td>-.12&lt;sup&gt;b&lt;/sup&gt;</td>
<td>.04</td>
<td></td>
<td></td>
<td>Comparison of unequal distances, Rule II</td>
</tr>
<tr>
<td>a</td>
<td>.86</td>
<td>1.29</td>
<td>.91</td>
<td>.92</td>
<td></td>
<td></td>
<td>Deciding whether the weight and distance dimension agree</td>
</tr>
<tr>
<td>c&lt;sub&gt;0&lt;/sub&gt;</td>
<td>2.39&lt;sup&gt;c&lt;/sup&gt;</td>
<td>-.53&lt;sup&gt;b&lt;/sup&gt;</td>
<td>-1.25</td>
<td>-.84</td>
<td>-.77&lt;sup&gt;b&lt;/sup&gt;</td>
<td></td>
<td>Compensation rule</td>
</tr>
<tr>
<td>b&lt;sub&gt;1&lt;/sub&gt;</td>
<td>1.00</td>
<td>1.13</td>
<td>.98</td>
<td>2.50</td>
<td>-.09</td>
<td></td>
<td>Number of buggies; compensation-rule</td>
</tr>
<tr>
<td>b&lt;sub&gt;2&lt;/sub&gt;</td>
<td>.52</td>
<td>.23</td>
<td>.24</td>
<td>.51</td>
<td>-.02&lt;sup&gt;b&lt;/sup&gt;</td>
<td></td>
<td>Maximum number of weights, compensation-rule</td>
</tr>
<tr>
<td>s&lt;sub&gt;1&lt;/sub&gt;</td>
<td>-.12&lt;sup&gt;b&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Sum of sums of weights and distance, compensation-rule</td>
</tr>
<tr>
<td>s&lt;sub&gt;2&lt;/sub&gt;</td>
<td>-.09&lt;sup&gt;b&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Symmetries, compensation-rule</td>
</tr>
<tr>
<td>s&lt;sub&gt;3&lt;/sub&gt;</td>
<td>-.29&lt;sup&gt;b&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Sum difference, compensation-rule</td>
</tr>
<tr>
<td>p&lt;sub&gt;0&lt;/sub&gt;</td>
<td>2.62&lt;sup&gt;c&lt;/sup&gt;</td>
<td>.27&lt;sup&gt;b&lt;/sup&gt;</td>
<td>-.22&lt;sup&gt;b&lt;/sup&gt;</td>
<td>.41&lt;sup&gt;b&lt;/sup&gt;</td>
<td>.57&lt;sup&gt;b&lt;/sup&gt;</td>
<td></td>
<td>Rule IV</td>
</tr>
<tr>
<td>p&lt;sub&gt;1&lt;/sub&gt;</td>
<td>-.76</td>
<td>-.37</td>
<td>-.16&lt;sup&gt;b&lt;/sup&gt;</td>
<td>.03&lt;sup&gt;b&lt;/sup&gt;</td>
<td>-.01&lt;sup&gt;b&lt;/sup&gt;</td>
<td></td>
<td>Symmetries, Rule IV</td>
</tr>
<tr>
<td>p&lt;sub&gt;2&lt;/sub&gt;</td>
<td>.12</td>
<td>.12</td>
<td>.15</td>
<td>.48</td>
<td>-.02</td>
<td></td>
<td>Product difficulty, Rule IV</td>
</tr>
<tr>
<td>p&lt;sub&gt;3&lt;/sub&gt;</td>
<td>-.03&lt;sup&gt;b&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Product difference, Rule IV</td>
</tr>
<tr>
<td>Parameter</td>
<td>Value</td>
<td>Rule/Description</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----------</td>
<td>-------</td>
<td>-----------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$g_0$</td>
<td>2.28c</td>
<td>Sum of weight and distance, Rule III</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$g_1$</td>
<td>.28</td>
<td>Hesitation parameter, Rule IV</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$p_4$</td>
<td>.74</td>
<td>Base of law of practice ($i_1 e^{-i_2}$)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$i_1$</td>
<td>1.83</td>
<td>Exponent of law of practice ($i_1 e^{-i_2}$)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$i_2$</td>
<td>.07</td>
<td>Inconsistency, Rule I</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>incr</td>
<td>2.67</td>
<td>Inconsistency, Rule II</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>incI</td>
<td>1.25</td>
<td>Inconsistency, Rule III</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>incII</td>
<td>1.23</td>
<td>Inconsistency, Rule IV</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>incIV</td>
<td>.68</td>
<td>Inconsistency, compensation-rule</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>inc_comp</td>
<td>1.35</td>
<td>Age</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$Y_1$</td>
<td>−.14</td>
<td>Age</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note.* All parameters indexed with superscript “0” are intercepts, whereas all parameters indexed differently are slopes. All parameter values are significant ($\alpha = .05$), except for the values indicated with superscript 0.

*a* Model was applied to the level of rules.

*b* Model was applied to the individual level.

*c* The estimated values are for parameters $w$, $d$, $c$, $p$, and $g$, respectively.

---
conflicted. Participants who used a complex rule (Rule III, Rule IV, or the compensation-rule) made this decision after they decided that both the values on the weight dimension and the values on the distance dimension differed.

Compensation. The extra time that was spent on solving conflict items with the compensation-rule was expressed as \( c_0 + b_1B + b_2 \max(W) + s_1S + s_2\Delta S + s_3\text{Sym} \). The estimate of the intercept of the expression, \( c_0 \), was \(-.53, t = -.87, p = .38\). Parameter \( b_1 \) was estimated at \( 1.00, t = 4.00, p < .01 \), implying that executing one buggy took \( 1.00 \) s. The estimate of \( b_2 \), associated with the largest number of weights in the two piles, was \(.52, t = 3.63, p < .01 \). Hence, increasing the number of weights in the pile that needed to be shifted with one increased RT by \(.52 \) s. Contrary to the buggy-parameters, the parameters associated with the addition-rule were estimated at non-significant values (\( s_1 \) was estimated at \(-.12, t = -1.73, p = .08 \); \( s_2 \) was estimated at \(-.09, t = -1.46, p = .65 \); \( s_3 \) was estimated at \(-.29, t = -1.88, p = .06 \)). The plots concerning the RTs of compensation-rule on conflict items in Fig. 4 clearly show that the addition of the buggy-parameters was impressively successful. The within item type variance on conflict items was almost completely predicted by model 2a.

Product comparison. The additional time that was needed to solve conflict items with Rule IV was expressed as \( p_0 + p_1\text{Sym} + p_2\sum P + p_3\Delta P \). The intercept, \( p_0 \), was estimated at \(.27, t = .70, p = .49 \). Parameter \( p_1 \) was estimated at \(-.76, t = -4.14, p < .01 \), showing that RT decreased if a conflict item contained a symmetry. Parameter \( p_2 \) was estimated at \(.12, t = 7.38, p < .01 \), demonstrating that an increase of product difficulty lengthened RT. Parameter \( p_3 \), associated with product difference, was estimated at the non-significant value of \(-.03, t = -.53, p = .60 \). Fig. 4 shows that model 2a successfully predicted the within item type variance on conflict items for Rule IV.

Rule III. Several hypotheses concerning the mediating effects of quantitative item characteristics on RT were tested for Rule III, but only the estimated value of the parameter that was related to the sum of the weight difference and the distance difference was significant. The added term in model 2, expressing the extra time that participants who used Rule III used for solving conflict items, was \( g_1(\Delta W + \Delta D) \). An intercept is missing because a model that contains both parameter \( g_0 \) and parameters \( p_0 \) and \( c_0 \) is not identified and hence prohibits estimation of model 2. Parameter \( g_1 \) was estimated at \(.28, t = 3.01, p < .01 \), showing that participants who used Rule III needed more time when both the values on the weight dimension and the values on the distance dimension showed large differences. Fig. 4 shows that the within item type variation on conflict items was not explained for Rule III.

Law of practice. Parameters \( i_1 \) and \( i_2 \), which were used to model the speed that participants gain during the test, were estimated at \( 1.69, t = 12.91, p < .01 \), and \(.06, t = 5.33, p < .01 \).

Sources of misfit. Fig. 4 shows that the predicted values for model 2a, represented with dashed lines, and the observed data, represented with continuous lines, were very similar. The additional parameters of model 2a certainly improved the description of the observed data considerably, compared to model 1. Clearly, the parameters that were introduced for weight comparison, distance comparison, guessing,
compensation, and product comparison had a great explanatory value. However, a number of problems remained, as is obvious from Fig. 4.

First, model 2a underestimated the observed RTs on simple-weight items for Rule IV, but overestimated the RTs on simple-weight items for all other rules. A possibility, which came up after unsuccessfully trying several other possibilities, was that users of Rule IV hesitated giving the response “scale tips to the side with the larger number of weights,” perhaps because they thought this response too trivial. Equation \( p_4 H \) was added in ad hoc model 3, where \( H \) was 1 when the correct response was “tips to the side with the larger number weights” and 0 in other cases. This addition changed the estimated values of the other parameters that were associated with the process of comparing unequal numbers of weights (\( w_0^x \) and \( w_1^x \)) as well. Initially, the RTs of all rules were modeled with only these two parameters, including the high RTs of Rule IV. As these RTs were additionally modeled with \( p_4 \), the estimates for parameters \( w_0^x \) and \( w_1^x \) lowered. Equation \( p_4 H \) did not agree with Siegler’s (1981) theory of the rule models and was therefore placed in curly brackets in Fig. 1. Parameter \( p_4 \) was estimated at .74, \( t = 5.34, p < .01 \). Including this parameter solved the problems with the simple-weight items to a large degree.

A second, more serious, source of misfit was the underestimation of the RTs on conflict items, especially on conflict-distance items, for Rule II. Possibly participants who used Rule II did not only consider the distances at which weights were placed when the numbers of weights were equal, but also when the numbers did differ. In model 3, it was assumed that Rule II-users compared the distances at which the weights were placed, whether the numbers of weights were equal or not. The mediating effects of the absolute values on the distance dimension and of distance difference were perhaps different for Rule II-users. Hence, an additional set of distance comparison parameters was formulated. The set for Rule II consisted of the equations \( d_{2,0}^- + d_{2,1}^- D \) (comparing equal distances) and \( d_{2,0}^x + d_{2,1}^x \Delta D \) (comparing unequal distances). The set for Rule III, the compensation-rule, and Rule IV still consisted of the equations \( d_{1,0}^- + d_{1,1}^- D \) (comparing equal distances) and \( d_{1,0}^x + d_{1,1}^x \Delta D \) (comparing unequal distances). In Fig. 1, the diamond that was associated with comparing distances on items with unequal numbers of weights was drawn with dashed lines for Rule II. The equation that referred to this process was placed between curly brackets to indicate that it did not agree with Siegler’s (1981) description of Rule II.

Fig. 5 shows the observed RTs of users of Rule II only, on all item types, in continuous lines. Dashed lines represent the predicted values of model 2a, whereas dotted lines represent the predicted values according to model 3. The latter values are much more similar to the observed RTs than the former, especially on conflict-distance items. The significant improvement in fit of model 3, compared to model 2a, \( F(5,327) = 20.21, p < .01 \), and the increase in explained variance to .88 demonstrated that the additional parameters, \( d_{2,0}^x, d_{2,1}^x, d_{2,0}^x, d_{2,1}^x, \) and \( p_4 \), improved the model.

Parameters \( d_{2,0}^- \) and \( d_{2,1}^- \) were estimated at .70, \( t = 1.96, p = .05 \), and \( -66, t = -2.53, p < .05 \), indicating that the influence of the absolute distances at items with equal distances was equal for users of Rule II and for users of complex rules:
RT decreased when the piles were placed at the pegs next to the fulcrum or at the furthest pegs.

Parameters $d_{2.0}$ and $d_{2.1}$ were estimated at .04 ($t = 2.23, p = .023$) and .34 ($t = 3.87, p < .01$). The mediating parameter for the effect of distance difference, $d_{2.1}$, was estimated at a positive value. A negative value of the parameter (like for the complex rules) implies that a large distance difference shortened RT. A positive value of the parameter (as for Rule II) implies that a large distance difference lengthened RT. So, a large distance difference facilitated an item for users of complex rules, but drew the attention of users of Rule II and hence lengthened their RTs.

The predicted values of model 3 were averaged per item type and represented with dashed lines in Fig. 3. These values approached the pattern of observed RTs satisfactorily and demonstrated that RTs can be predicted quite precisely at the level of rules.

**Analyses of RTs on the individual level**

The individual level of analysis concerned the RTs of each individual and regarded 182 (participants) * 70 (items) = 13,370 data points. In model 4, we applied the equations of model 3 to this level of analysis. Although the estimates of both models were quite similar, the explained variance of this model was only .16. Clearly, there was much more variation at the individual level.

**Age.** The factor “Age” was added in model 4a by means of a linear effect. Perhaps a non-linear effect can better model the factor age, but the gap in the distribution of age, roughly between 13 and 17 years, restrained us from an extensive analysis.
of age. The explained variance increased to .18 when the parameter associated with age, \( y_1 \), was added. An ANOVA comparison indicated that the improvement of model 4a was highly significant, \( F(1, 12, 646) = 354.06, p < .01 \), compared to model 4. The estimated value of \( y_1 \), was \(-.14, t = -18.82, p < .01\), indicating that an increase of one year was accompanied with a decrease in RT of .14 s.

**Rule inconsistency.** In model 4b, inconsistency of rule use was included. As described before, a participants’ response pattern was classified to a rule, based on the results of the cluster analysis. Each cluster had an expected response pattern (simply derived from the so-called cluster centers, see Table 4). For each participant and for each item, the absolute distance between the expected item score (based on the cluster the participant was classified in) and the participant’s item score was calculated. If, for instance, a Rule I-user succeeded on a simple balance item, the inconsistency score (\( \text{Inc} \)) was \([1.00 - .97] = .03\); when a Rule I-user unexpectedly failed a simple balance item, the inconsistency score was \([.00 - .97] = .97\).

Incorporating rule inconsistency improved the fit of the model significantly, \( F(5, 12, 641) = 22.00, p < .01 \), compared to model 4a. The explained variance increased to .19. Table 5 shows the estimates of model 4b. As expected, the effect of rule inconsistency differed between rules. The inconsistency parameter was 2.67 for Rule I, 1.25 for Rule II, 1.23 for Rule III, 1.60, and 1.35 for the compensation-rule, respectively. As expected, the estimate of the inconsistency parameter was largest for users of Rule I and not significant for users of Rule IV. Table 5 shows that the estimates of the remaining parameters were quite similar to those of previously fitted models.\(^4\)

**Age interaction.** Age interactions were introduced for all mediating factors in model 5 as all processes may happen faster for older participants. The estimates for the parameters of model 5 are printed in two columns in Table 5. The first column is organized like the columns for the other models, whereas the second column shows the estimated values for the interactions with age. A positive value implies that the process is executed slower with older participants, whereas a negative value implies that the process happens faster with older participants. The explained variance of model 5 was .20. The fit of the model was significantly better than the fit of model 4b, \( F(14, 12, 627) = 9.710, p < .01 \). Age interactions were significant for three parameters. The age interaction for parameter \( d^2_{11} \) was estimated at .04, implying that the comparison of unequal weights took longer for older participants when the distance difference was large. The age interaction for parameter \( b_1 \) was \(-.09\), indicating that older participants performed buggies faster than younger participants did. The age

---

\(^4\) With these estimates, the expected values for each subject on each item can be calculated. Consider a participant who is 14 years old. His predicted RT on the fifth item, a conflict-weight item with three weights on the second peg on the left, and one weight on the fourth peg on the right side of the scale, which he answers consistent with the compensation-rule, is:

\[
E(\text{RT}) = \left(w_4^e + w_1^e \Delta W + (d^2_{11} + d^1_{11} \Delta D) + a + (c_0 + b_1 B + b_2 W) + i_1 e^{-bi} + (y_1 A) + \text{inc}_{\text{comp}} \ast \text{Inc} \right) \\
\iff E(\text{RT}) = (4.08 - .05 \ast [3 - 1]) + (1.95 - .49 \ast [2 - 4]) + .91 + (-.84 + .98 \ast 2 + .24 \ast 3) + 1.76e^{-0.75} \\
+(1.35 \ast .00 - .15) + (-14 \ast 14) \iff E(\text{RT}) = 7.18s.
\]
interaction for parameter $p_2$ was estimated at $-.02$, again indicating that older participants needed less time to calculate large products than younger participants did.

**Mixed effects.** The explained variance of model 5 was still moderate, contrary to that of model 3 at the level of rules. Clearly, the individual variance was high. Individual differences can be further modeled by applying mixed effects models (Pinheiro & Bates, 2000). In model 6, a random effect for parameter $w_0^d$ was incorporated as this parameter serves as the basic intercept. The explained variance rose to .33 and the improvement of the model fit, as assessed with the likelihood ratio statistic, was highly significant, $LR(44) = 1724.39$, $p < .01$, compared to model 5. Adding a random effect for parameter $i_1$, and thereby allowing a random effect for the slope of the regression equation, in model 6a, improved the fit of the model even more, $LR(2) = 204.41$, $p < .01$, compared to model 6. The explained variance of model 6a was to .36. The standard deviation of the random effect for the intercept was 1.75, whereas the standard deviation for slope was .02. The correlation between the random effects was $-.178$.

**RTs on weight-distance items**

A total of 72 children responded to the six weight-distance items. This sample included 14 Rule I-users, 14 Rule II-users, 20 Rule III-users, 2 Rule IV-users, 18 users of the addition-rule, and 4 participants who were using an unknown strategy. These last four participants were excluded from the analyses. The remaining participants were classified as using either a simple (Rule I- and Rule II-users) or a complex rule (Rule III, the compensation-rule, and Rule IV). The difference between the average age of participants using a simple rule (10.6 years) and the average age of participants using a complex rule (12.1 years) was significant, $t(64) = -4.97$, $p < .01$. The observed RTs clearly supported the hypothesis that participants who used a simple rule responded faster to weight-distance items than participants who used a complex rule. The difference between the short RT of the young children using a simple rule ($M$ (RT) $= 2.63$) and the long RT of the older children using a complex rule ($M$ (RT) $= 3.86$) was significant, $t(52) = -5.02$, $p < .01$, supporting our claim that older children were slower than young children, even on items that were solved correctly by all participants.

The continuous lines in Fig. 6 show the observed average RTs on weight-distance items for the simple rules (left panel) and the complex rules (right panel). From the basic model of Siegler (1981, see Table 2), it was predicted that the RT on weight-distance items for participants who used a simple rule only consisted of comparing weights ($w$). The RT for participants who used a complex rule was expected to consist of the time that it took to compare weights, to compare distances, and to decide that the weight dimension and the distance dimension agreed ($w + d + a$).

The predictions that followed from model 3 were more complicated. For Rule I, it was expected that RT consisted only of the time that was spent on comparing unequal weights, which was influenced by the difference between the number of weights on the left and the number of weights on the right side of the scale ($w_0^d + w_1^d \Delta W$). Participants who used Rule II were also expected to compare distances as well
Participants who used a complex rule were expected to execute the extra step that was needed to decide that the dimensions agreed ($a$). The RTs on weight-distance items also depended on the timing of the items, for all rules.

Dashed lines in Fig. 6 represent the predicted values according to the model 2a, whereas dotted lines represent model 3. The predictions fit the observed RTs reasonably well. Especially the within item type variation is predicted quite satisfactory. The predictions are certainly acceptable since the parameters that determined the predictions originated from a model that was not optimized on the observed RTs on the weight-distance items. Model 2a predicts the wd data even better than model 3. In view of the limitations of the weight-distance data (all weight-distance items at the end of test for only a subset of the participants), we will not try to interpret this difference in prediction.

The RTs on weight-distance items provided a second way of estimating parameter $a$. Parameter $a$ was estimated by subtracting the average RT of users of complex
rules on simple items from the average RTs of these users on weight-distance items because the RT on weight-distance items (3.86 s) was expected to consist of the same processes \((w + d)\) as the RT on simple items (3.03 s) and the additional process of deciding whether the dimensions agreed \((a)\). Hence, parameter \(a\) was estimated at .83 s.

**Discussion**

*Rule use on the computerized balance scale task.* The classification of participants’ responses on the computerized balance scale task corresponded largely to the classification of participants’ responses to individually administered balance scale tasks (e.g., Siegler, 1981) and the classification of participants’ responses to group-wise administered paper-and-pencil versions of the balance scale task (e.g., Jansen & van der Maas, 2002; Van Maanen et al., 1989).

*RTs on the balance scale task.* The RTs on the balance scale items were modeled by means of regression models. First, the predictions that followed from the Siegler’s (1981) model (model 1) were tested. This basic model consists of a hierarchy of increasingly complex rules. The rules consist of general steps, like weight comparison and distance comparison, and of rule-specific steps, like calculating and comparing products of weights and distance. Fitting a regression model resulted in estimates for the stated steps. Although the RTs, inferred from the basic model, approached the observed RTs reasonably, the model was refined considerably. The alterations resulted in model 3, which is depicted in Fig. 1.

In the basic model, participants who used Rule I were expected to compare the numbers of weights on the scale on all balance scale items. The time that was needed to compare equal numbers of weights was predicted to equal the time that was needed to compare unequal numbers of weights. However, model 3 showed that the comparisons corresponded to different processes. Deciding that two piles contained the same amount of weights took more time when the piles contained a large number of weights. The duration of the basic process of comparing equal weights was 2.02 s and the increase of one weight in each pile caused an increase in RT of .18 s. It was tested whether deciding that the numbers of weights in two piles differed was affected by the difference between the numbers of weights in the piles. This was not the case; a difference between the numbers of weights was easily detected, in 2.60 s, independent of the size of the difference.

Participants who used Rule II were also expected to compare the numbers of weights. Moreover, they were expected to compare the distances at which the weights were placed when the numbers of weights were equal (i.e., on simple-balance and on simple-distance items). From model 3, it was concluded that Rule II-users compared distances at any item, whether the numbers of weights were equal or not. Therefore Fig. 1 contains an extra step for Rule II, the diamond drawn with dashed lines. The response patterns showed that Rule II-users always answered that the scale would tip to the side with the larger number of weights, but their RTs showed that they did consider the distance dimension when the numbers of weights differed.
It was concluded that deciding that the piles of weight were placed at the same distance from the fulcrum was hindered by the position of the piles. Deciding that the distances were identical was easy 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 .34 s. Perhaps they were attracted to the distance dimension when the difference was large and spent more time on it than when the difference was small and hence less salient. Older users of Rule II were even slower on comparing unequal distances when distance difference increased than younger users of Rule I.

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 .41 s. Probably, the difference between the distances was more easily noted when the difference was large.

Users of complex rules needed to judge whether the distance dimension and the weight dimension agreed when values on both dimensions differed. This process was estimated to last 1.29 s. Parameter \( a \) was also estimated by subtracting the RTs of users of complex rules on simple items from the RTs of these users on weight-distance items. This resulted in an estimation of .83 s. Parameter \( a \) was probably overestimated in the model because it included a guess-process as well. A model that contained both a guess parameter and parameter \( a \) was not identified. However, the procedure of subtracting RTs on simple items from RTs on weight-distance items was unsophisticated as it did not involve separate parameters for the comparison of equal or unequal values, quantitative attributes or timing of the items.

It turned out to be very difficult to explain the variation in RT on conflict items for participants who used Rule III. It was concluded that the RT of these participants was .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. The response "scale remains in balance" was very rare among the response patterns of the participants who were classified in Rule III. It was concluded that they did not choose from the three response options, but always thought that one dimension dominated the other and that dimensions never compensated each other. It remains unknown how these participants decided which dimension was more important. However, it seems likely that this 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 of the compensation-rule and Rule IV.

Considering the response patterns can not differentiate the addition-rule from the buggy-rule. The predictions concerning the RTs that were associated with the rules
did differ. We conclude that participants used the buggy-rule and not the addition-rule because model 3 demonstrated that RT was affected only by parameters that were related to predictions that followed from the buggy-rule. The buggy-rule is used on conflict items and implies that participants shift the pile with the larger number of weights (and the smallest distance from the fulcrum) towards the end of the scale. For each shift (“buggy”), they take one weight off the pile. Shifting the pile ends when the numbers of weights or when the distances are equal. The RT of participants who used the compensation-rule increased with 1.13 s for each required buggy. As children grow older, they become faster at performing buggies. Next, the RTs associated with the compensation-rule increased with .23 s for each weight in the pile that needed to be shifted.

It should be noted that the response pattern of the cluster that was associated with the compensation-rule did show deviations from the theoretically expected response pattern. Possibly, the computer test, and the presence of the test assistant, elicited faster responses from participants than a paper-and-pencil test does. In this study, participants responded to over 70 items in, on average, 18 min, which is very fast. The fast responding may have prevented participants from carefully executing the compensation-rule.

Model 3 showed that Rule IV does not only involve executing the torque rule on conflict items, but also considering whether a conflict item contains any symmetries: RT was .37 s shorter when the distance at which the weights were placed on one side of the scale was equal to the number of weights on the other side of the scale. Moreover, RT was lengthened with .12 s when the sum of products of weight and distance increased with one, resulting in an increase of product difficulty. Age interactions showed that older participants were faster at calculating products.

An extra parameter was added to improve the description of the RTs of users of Rule IV. This parameter referred to the hesitation, lasting .74 s, to answer that the scale would tip to the side with the larger number of weights. This process explained the long RTs of users of Rule IV on simple-weight items. Parameter $p_4$ did not agree with the theoretical rule model and was therefore placed between curly brackets in Fig. 1. Although the parameter did improve the fit of the model, we do not consider it a very reliable parameter.

Performance on weight-distance items. Weight-distance items include a larger number of weights on one side of the scale, placed at a greater distance than the smaller number of weights on the other side. Siegler’s (1981) rule description resulted in the prediction that participants, who used Rule I or Rule II, immediately decided, after observing that the numbers of weights differed, that the scale would tip to the side with the larger number of weights. On the contrary, it was expected that participants who used Rule III, the compensation-rule, or Rule IV, executed two additional steps before they derived the correct answer. The first involved the assessment of the distance dimension, whereas the second involved the decision that the larger number of weights was on the same side as the larger distance. The predictions were refined in model 3. It was expected that the duration of the steps depended on the weight difference, distance difference, and the timing of the item. Moreover, it was expected that participants who used Rule II also compared distances at weight-
distance items. Nevertheless, the main hypothesis, that users of simple rules were faster on weight-distance items that users of complex rules, was confirmed.

Rule inconsistency. Assessment of rule use was based on a cluster analysis of the responses on all 70 items in the test. Differences between the observed response patterns and the cluster pattern suggested that a rule was not always performed consistently. A positive relation was observed between rule inconsistency and increase of RT. The effect of rule inconsistency on RT was largest for users of Rule I and absent for users of Rule IV.

Conclusions

The above results showed that the analysis of RTs revealed meaningful information on the cognitive processes underlying problem-solving behavior on the balance scale task. The results provided evidence for Siegler’s (1981) rule models of the balance scale task and demonstrated that young adults were indeed slower than children when solving balance scale problems. Perhaps young adults trade speed for accuracy (speed-accuracy trade-off phenomenon; Luce, 1986) but the outcome of the fitted models suggested that it was more likely that the complexity of the rules that young adults used for solving balance scale items caused their lengthy RTs.

The observed RTs, at the level of rules, were predicted almost perfectly with the basic steps, introduced by Siegler (1981), and newly introduced parameters associated with quantitative features of the items. The fit of the eventual model was convincing because the proportion of explained variance reached the remarkably high value of .88.

The first important conclusion was that participants who used Rule II did not only consider the distance dimension when the values on the weight dimension were equal. They seemed to consider the distance dimension at all items, but without knowing how to combine it with the weight dimension, resulting in the dominant response “scale tips to the side with the largest number of weights.” Only when the values on the weight dimension were equal did they incorporate the distance dimension in their response. This matches Siegler’s (1981) observation that many children whose response patterns indicated the use of Rule II said that they did consider both the weight and the distance dimension on each balance scale item. Older participants who used Rule II were even slower on items with unequal distances than younger participants who used Rule II. Possibly, the transition from Rule II (noticing distance but only using it when the values on the weight dimension are equal) to Rule III (guessing when both the values on the weight dimension and the values on the distance dimension conflict) happens gradually (Jansen & van der Maas, 2002). In this case, it is possible that, as children grow older, they pay more and more attention to the distance dimension until they incorporate it in their solution strategy.

The second important conclusion was that participants did not use the addition-rule but were using the buggy-rule. These rules theoretically produce the same response pattern. However, the RTs associated with this response pattern were affected only by variations that, theoretically, mediated the performance of the buggy-rule. It
should be noted that a possible effect of the performance of an addition-rule, the difficulty of the sums that need to be calculated, was maybe too small to be noted. The average age of participants who used the addition-rule was 16 years and the addition of single-digits is easy for participants of this age as the solution for sums of small digits is often retrieved (Siegler, 1998).

The third conclusion concerned the effect of the exact design of an item on the duration of the cognitive processes that made up the rules. First, the duration of comparing equal values was different from the duration of comparing different values on a dimension. Second, the exact values of the dimensions influenced the RT. Generally, the higher the values, the more time it took to inspect an item. However, when the difference between values was large, RT decreased, in general. We concluded that the RTs did not show homogeneity for items of one type, although response patterns did: the rule with which a participant approached an item was equivalent for all items of one type, but the duration of the execution of the rule depended on the item presented.

The rule-specific relation between rule inconsistency and increased RT provides additional, although moderate, evidence for Jansen and van der Maas’ (2002) developmental model of reasoning on the balance scale task. The model assumes a discontinuous shift from Rule I to Rule II, which was modeled with the cusp model (Jansen & van der Maas, 2001). This model predicts, in cases of discontinuity, an increased RT after a switch has taken place: “critical slowing down.” The finding that rule inconsistency was associated with increased RT for Rule I and Rule II was in accordance with this prediction and provided additional evidence for the hypothesis that the shift from Rule I to Rule II happens discontinuously. Also in line with expectations was the finding that rule inconsistency did not affect the RTs of participants using Rule IV. Once participants have mastered this rule, there is no need to switch to another rule as Rule IV always results in the correct response.

As rule inconsistency can have several causes, it would be interesting to study each participant’s pattern of inconsistency. Rule switches would cause rule inconsistency on specific item types only. For instance, rule switching between Rule II and the buggy-rule may result in inconsistency on conflict items but not on simple items. The most interesting case of rule switches are shifts to a rule that is higher in the hierarchy. In this case, rule inconsistency will only be observed before or after the switch (depending on whether the simple or the more complex rule is used more often). Erring, caused by boredom or fatigue of performing a rule, can also result in inconsistency and cause deviating RTs. More intensive study of rule inconsistency, and the associated increased RTs, is needed.

Although the expanded Siegler model described the RTs very well, the model fit at the individual level was modest. It is interesting to note that the explained variance differed considerably over rules (.10, .16, .21, .37, and .31 for Rules I, II, III, IV, and the compensation-rule, respectively). Adding random effects at the individual level improved the fit considerably, which indicates that variation was high at the individual level. Part of the variability was caused by the presence of outliers. Other causes of variability are yet unknown.

In future research, the test needs to be modified in several ways. In the present test, all conflict-distance items and conflict-balance A items could be solved correctly.
by means of the compensation-rule, whereas all conflict-weight and conflict-balance B items could not be solved correctly by using this rule. This was done to ease the detection of the compensation-rule. However, this connection obstructed testing some hypotheses regarding RTs, like the influence of symmetries on the duration of the compensation-rule.

Furthermore, some quantitative characteristics were equal for all items of one type but differed from those of other item types. For example, all conflict-distance items contained the same, maximum value, of distance difference. This sometimes made it impossible to test whether variations in RTs were caused by the quantitative characteristic concerned or by the specific item type. In future research, quantitative characteristics should be independent of item type.

Some quantitative characteristics of items were highly correlated. This should be avoided in a future test, although the problem can never be ruled out completely. For instance, sum difference and product difference are necessarily correlated.

Finally, possible effects of the sequence of item types were not tested. In future research, the sequence may be randomized. The reported results concerning weight-distance items were promising and warrant further research as weight-distance items make it possible to test Siegler’s (1981) rule models more accurately. Weight-distance items should be placed throughout the test, instead of at the end of the test, as was done in the present study.

Designing items that tap each proposed step separately would meet the general comments on the subtraction method that we used here. The assumption of the possibility of splitting the RT of a complex process into RTs of simple processes can be tested if both the RTs, associated with the simple processes, and the RT, associated with the (combined) complex process, are known. Also the assumption of “pure insertion” can be tested with these kind of data.

In spite of these difficulties, the findings in this paper are promising and demonstrate that the analysis of RTs, in addition to response patterns, is very informative. Although the response patterns are most useful to detect the rule that a participant employs, the RTs provide cross-validation of the proposed contents of the rules and even present new information on the processes that participants employ when solving balance scale problems.

The analysis of RTs may also be applied in other domains of cognitive development. This study demonstrates that analysis of RTs in the range of 2–10s are very informative on cognitive processes and on cognitive development. The results show that RTs are an important part of an individual’s profile of performance and that RTs can add considerably to our understanding of the cognitive processes that underlie problem solving. Finally, the richness of the present data constitute a clear challenge for the many computational models proposed for the balance scale task.

References


