Development of Rules and Strategies: Balancing the Old and the New

Robert S. Siegler

Carnegie Mellon University

and

Zhe Chen

University of California, Davis

The experiments described in the lead articles replicate findings from previous studies of development of knowledge about balance scales, add several new findings, and raise four key questions: (a) How can rule use best be assessed? (b) How can we reconcile systematic use of rules with variable use of strategies? (c) When do children begin to use rules? and (d) How do children generate new rules? In this Reflection, we summarize current understanding of development of knowledge about balance scales and consider each of the four questions. © 2002 Elsevier Science (USA)

Key Words: cognitive development; scientific reasoning; balance scales; rule use; rule assessment; strategies; encoding.

One of the most gratifying experiences a scientist can have is seeing that many years after publishing a set of studies, the studies continue to inspire important new research. Thus, it was with special interest that we read the lead articles in this issue. The new findings extend in interesting ways previous understanding of development of knowledge about balance scales and also illustrate the usefulness of a promising new method for addressing some of the unresolved issues in the area. In this Reflection, we consider what was known about development of balance scale expertise before these articles were published, what the new findings add, and what we need to learn to advance beyond current understanding.

The balance scale task has become a benchmark for testing the usefulness of alternative methodologies and theoretical approaches. Among the methods that have been used to examine the development of understanding of balance scales are protocol analysis (Inhelder & Piaget, 1955/1958), rule assessment (Siegler,
1976), information integration (Surber & Gazesh, 1984), item-response analysis (Wilson, 1989), and now latent class analysis (Jansen & van der Maas, 2002, this issue). Among the computational approaches used to model the acquisition process are production systems (Klahr & Siegler, 1978; Langley, 1987), parallel distributed processing networks (McClelland, 1995; Raijmakers, van Koten, & Molenaar, 1996; Shultz, Schmidt, Buckingham, & Mareschal, 1995), SOAR (Newell, 1990), and ACT-R (van Rijn, 2001). Thus, even before publication of these new articles, there was a substantial body of knowledge regarding the development of balance scale expertise.

Despite the diverse methods that have been used to study children's understanding of balance scales, the research findings have been impressively consistent. Among the recurring findings of past research (e.g., Siegler, 1976, 1981), many of which receive further support in the current articles, are the following:

1. The large majority of 5-year-olds and older children consistently use rules to solve balance scale problems. Patterns of correct answers and errors, explanations of reasoning, and the agreement between error patterns and explanations all support this conclusion.

2. Children age 3 years or under rarely apply rules systematically to solve balance scale problems. Thus, rule use increases substantially between ages 3 and 5 years.

3. Balance scale rules are ordered hierarchically. Children start out basing predictions only on weight (Rule I). Then they base predictions on weight and, when weights are equal, on distance (Rule II). Then they predict in all circumstances on the basis of both weight and distance but do not know how to combine them quantitatively (Rule III), so they muddle through, guess, or use imperfect integrative approaches such as the addition rule, in which they add the weight and distance on each side of the fulcrum and choose the side with the greater sum. Finally, some adolescents and adults (but only a minority) learn the torque rule, which consistently solves problems correctly (Rule IV).

4. Rule use is correlated with age. The large majority of 5- and 6-year-olds rely on Rule I. Some 8- and 9-year-olds use Rule I, others use Rule II, and yet others use Rule III. Most 13-year-olds use Rule III or the addition rule. Finally, the majority of older adolescents and adults use Rule III or the addition rule, and a minority of these individuals use Rule IV.

5. Extreme variations in distance from the fulcrum can lead 5- and 6-year-olds who typically base responses only on weight to consider distance as well.

6. Feedback can help children to learn more advanced approaches than those they use in the absence of feedback. Children's likelihood of learning more advanced rules from feedback is greatly influenced by their encoding of dimensions that are not yet part of their predictive rule. Thus, children who use Rule I but who also encode the distance dimension more often learn Rule II from feedback problems than do children who also use Rule I but who do not encode the distance dimension.
These previous findings provide the background for considering the contributions of the lead articles to our understanding of four questions.

**How Can Rule Use Best Be Assessed?**

The rule assessment method (Siegler, 1976, 1981) defines rule use in terms of an individual’s consistent application of an algorithm across a large and diverse set of problems. It assumes that different children of a single age can and often do use different rules, but that an individual child adheres to a single rule during a session. Consistent with this emphasis on an individual child’s consistent application of a single approach, studies using the rule assessment method have adopted quite conservative criteria for inferring rule use. The typical criterion for concluding that a child used a given rule was that at least 80% of the child’s responses on a large number of three-choice problems were the response predicted by the rule. Often, a child also had to meet additional criteria to be said to have used a rule. For example, in cases where only a small number of problems discriminated between two rules, a child would need to predict in accord with one of the rules on at least 75% of those items, as well as on more than 80% of all items, in order to be said to have used the rule.

The rule assessment method, like any method, has both advantages and disadvantages. One advantage is that, due to the conservative criteria for rule use and the large and diverse sets of problems over which the rule must be applied, the statement that a given child used a given rule is a strong statement about the child’s behavior. When the large majority of children of a given age meet the criteria for using a particular rule, the finding indicates substantial regularity in the behavior of the entire age group.

A related pair of strengths of the rule assessment approach are that its results are stable over time and that they converge with other measures of knowledge. Evidence for the stability over time of rule use comes from Siegler (1981), in which the balance scale rule use of 3- to 9-year-olds was tested on two occasions a month apart in time. The large majority of children were classified as using the same rule on the two occasions. In the same study, when the children were asked to explain how they solved the balance scale problems, their verbal reports quite consistently were in agreement with the rule assessment that was derived from their pattern of predictions. The one exception was that many children whose predictive performance indicated that they used Rule II said that they relied on weight and distance without indicating that they relied only on distance in the limited case where weight was equal on the two sides of the fulcrum. Thus, the predictions and explanations data converged, but the rule assessment approach yielded a more precise depiction of when the children used the distance information.

A fourth strength of the rule assessment approach is that it has revealed considerable consistency of reasoning across problems. Thus, 5-year-olds not only use rules based on a single dimension on balance scale problems, they also use parallel unidimensional rules on shadow projection, liquid quantity conservation, and solid quantity conservation problems (Siegler, 1981); on fullness problems
A fifth strength of the rule assessment approach is its usefulness for predicting learning. By indicating the types of problems that cannot be solved using a given rule but that can be solved using its successor, the approach suggests what types of problems children should be able to learn most readily. For example, on the balance scale task, distance problems can be solved by Rule II but not by Rule I; such problems proved much easier to learn than other problems that also could not be solved by using Rule I (Siegler, 1976, Experiment 2).

A sixth strength of the rule assessment approach is that it can point to processes that are crucial for learning more advanced rules. Again using the transition from Rule I to Rule II on the balance scale as an example, the rule assessment approach, along with informal observation of children as they performed the task, suggested that failure to encode the distance dimension might be impeding adoption of more advanced rules. This hypothesis was tested and proved consistent with subsequent data (Siegler, 1976, Experiment 3).

Jansen and van der Maas (2002, this issue) identify several important advantages of latent class analysis that they claim differentiate it from the rule assessment method. One advantage that they cite is that latent class analysis offers measures of the statistical fit between the data and each rule model. A second advantage is that it avoids basing decisions of rule use on arbitrary criteria. A third is that it allows some deviations from the predictions of the rule. A fourth is that it allows falsification of hypothesized rules. A fifth is that it can lead to discovery of new rules by identifying clusters of unexpected response patterns.

We agree with Jansen and van der Maas that these are advantages of latent class analysis. We also agree with them that the introduction of measures of the statistical fit between the data and each rule model is an advantage of latent class analysis relative to the rule assessment approach. However, whether the other qualities differentiate the two approaches is far from clear. Are statistical criteria, such as that the log-likelihood ratio exceeds the .05 probability level, less arbitrary than the criterion that 80% of responses to large numbers of three-choice items need to be in accord with a given rule? Both are arbitrary conventions; just as the 80% cutoff point could be 70% or 90%, so too could the .05 significance level be .01 or .10. The 80% criterion, like the statistical criteria within latent class analysis, allows some deviations from the predictions of the rule; the most common instantiation of the approach, 20 of 24 responses in accord with the rule, allows 4 deviations. Numerous rule assessment studies, such as Siegler and Richards’s (1979) study of time, speed, and distance concepts, have failed to find evidence that children’s predictions conformed to some of the expected patterns; these findings falsified the hypothesized rules in the same sense that latent class analysis can falsify hypothesized rules.

Finally, the rule assessment approach, like latent class analysis, allows discovery of new rules through identification of clusters of unexpected response patterns. In one study that illustrated this property, Siegler and Vago (1978) presented 5- to
10-year-olds a task on which they needed to judge the relative fullness of two beakers of water. Many children’s responses followed the predicted pattern of choosing as more full the glass with the taller liquid column or with the greater volume of water. However, a number of other children generated patterns of correct answers and errors that did not resemble any predicted pattern but that were similar across those children. Examination of their response patterns revealed that they were using an empty space rule; the glass that contained less empty space atop the liquid column was fuller than the other glass. (Predictions from this rule differed from predictions of the rule based on volume of water because glasses of a variety of sizes were used. Thus, a glass that had more liquid than another glass could have more, less, or equal amounts of empty space atop the liquid.) Ability to uncover unanticipated rules is indeed an important quality of latent class analysis, but it is also an important quality of rule assessment analysis.

In addition to sharing many of the advantages of latent class analysis, the rule assessment approach has several advantages that have not been shown to be shared by latent class analysis. Some of these advantages are empirical properties that have been demonstrated for the rule assessment approach and that have not been demonstrated for latent class analysis—but that may be demonstrated for it in the future. These properties include stability of rule classifications over time, convergence between the rule assessments and verbal explanations, and usefulness for understanding learning.

Other advantages of the rule assessment approach relative to latent class analysis seem to stem from inherent differences between the approaches. Rule assessments can be done with small samples, whereas latent class analysis requires very large samples. Also, rule assessments guarantee that the behavior of a child said to be using a rule is stable over diverse types of problems. By contrast, most latent class analyses are conducted on a single type of item, and it is impossible to know from them whether an individual uses the same approach across different problem types and trial blocks.

In an interesting attempt to surmount this limitation and extend the latent class approach, Jansen and van der Maas (2002, this issue) conduct an analysis in which they compare individual children’s performance on three problem types in Blocks 2 and 3 to the children’s performance on the same problem types in Blocks 4 and 5. The analysis reveals all four rules reported in Siegler (1976) and also the addition rule reported by Normandeau, Larivee, Roulin, and Longeot (1989), a “smaller distance” rule, and the Rule III/addition rule. As indicated in Table 7 of the Jansen and van der Maas article, however, the latent class analysis classifications are quite unstable from the earlier to the later blocks in whether they categorize children as using Rule II, Rule III, the addition rule, the smaller distance rule, or the Rule III/addition approach. For example, only 39% of children who used the Rule III/addition approach are classified as using it both on Blocks 2 and 3 and on Blocks 4 and 5. This instability is present despite the assessment being conducted under optimal conditions for revealing stability—a single relatively short session, no variation in instructions or conditions of test-
ing, no feedback, same experimenter, and so on. This instability calls into question the usefulness of latent class analysis as a technique for revealing stable rule use of individual participants.

Jansen and van der Maas propose that the instability might have been due to learning. However, the data do not support this interpretation. Downward movement in sophistication of rules was roughly as common as upward movement (e.g., as many children moved from Rule III on Blocks 2 and 3 to Rule II on Blocks 4 and 5 as made the opposite transition).

These results illustrate the complex trade-offs involved in deciding the level of precision at which rules should be characterized. Rule III, in which children are said not to adopt any systematic approach on items on which the greater weight and the greater distance are on opposite sides of the fulcrum, is a less precise and less satisfying representation of children’s knowledge than the other rules proposed in Siegler (1976). The Rule III representation was adopted despite these disadvantages because the children said to be using it did not appear to adhere consistently to any more specific rule over the full set of items. Application of the very powerful latent class analysis technique to the responses of more than 700 children allowed Jansen and van der Maas to identify more precise patterns within the Rule III umbrella category. However, these response tendencies were so unstable that they often could not be detected from one trial block to the next. This raises serious questions concerning whether it is worthwhile to characterize these unstable response tendencies as rules rather than as a less consistent form of responding that can vary from trial to trial. The overlapping waves approach was designed precisely to facilitate analysis of such cognitive variability.

How Can We Reconcile Systematic Use of Rules with Variable Use of Strategies?

Studies using the rule assessment approach (e.g., Siegler, 1976, 1981) emphasized the orderliness of children’s rule use and indicated that children often use a single rule on a task. This approach worked well for the balance scale and a number of other problems: liquid and solid quantity conservation; projection of shadows; probability estimation; fullness; and time, speed, and distance comparisons (Siegler, 1981; Siegler & Richards, 1979). The tasks on which the rule assessment approach worked well shared several characteristics. The task was unfamiliar to children. It required a quantitative comparison (e.g., more liquid, fuller, faster, bigger shadow). It included two or more relevant dimensions, one of which was much more perceptually salient than the other(s) (e.g., weight and distance from the fulcrum on the balance scale, size of object casting the shadow and its distance from the light source and the screen on projection of shadows, height and cross-sectional area on liquid quantity conservation). On such tasks, children usually followed systematic rules such as relying on the single, perceptually salient dimension to the exclusion of the other relevant dimensions.

Subsequent examination of a broader range of problems, including ones that children often encounter and that do not involve quantitative comparisons, indi-
cated that such systematic rule-governed responding was characteristic of only a minority of tasks (Siegler, 1996). These problems include arithmetic, spelling, word identification, time telling, locomotor activity, language use, moral reasoning, and serial recall. On these and many other tasks, children use different approaches on different trials rather than consistently relying on a single approach. Children who use a given approach on one trial frequently shift to another approach on the next trial. This occurs even when children answer correctly on the first problem and the two problems are similar. Even when children were presented the same arithmetic or time-telling problem by the same experimenter under essentially identical conditions on 2 days within a 1-week period, one-third of children used different strategies on the two occasions (Siegler & McGilly, 1989; Siegler & Shrager, 1984). This variability of strategy use could not be explained by learning. Children in the time-telling and arithmetic experiments were nearly as likely to shift from more advanced strategies to less advanced ones as to shift in the opposite direction.

Such findings led Siegler (1996) to propose the overlapping waves model as a way of representing both the more common case, in which children know and use multiple strategies, and the less common case, in which they consistently employ a single rule. The overlapping waves model's basic assumptions are that individual children typically know and use multiple strategies; that the strategies usually coexist and compete with each other over prolonged periods of time; and that with age and experience, children increase their use of the relatively advanced approaches and also discover new strategies. To avoid confusion about the meaning of terms, Siegler labeled goal-directed approaches that varied from trial to trial as "strategies" and labeled goal-directed approaches that were applied consistently over trials as "rules."

Although Jansen and van der Maas state explicitly that they use the terms rule and strategy interchangeably, they eventually arrive at a similar distinction. In the intriguing proposal depicted in the bottom panel of their Fig. 2, they suggest that even development of a single task can include prolonged periods in which children respond in systematic, rule-governed ways and other prolonged periods in which children use multiple strategies and vary their approach from trial to trial. The bottom panel of their Fig. 2 illustrates how development on the balance scale may involve both types of development. It also illustrates one of the attractive qualities of the overlapping waves depiction: its ability to represent fast and slow transitions within the same framework. (The overlapping waves approach assumes that most transitions are gradual, but there is nothing in it that says that all must be.)

Although Jansen and van der Maas's proposal regarding development on the balance scale is intriguing, it also goes well beyond the data. Without longitudinal observations, it is impossible to know how abrupt balance scale transitions actually are. Their Fig. 2 depiction is based in large part on the number of switches in latent class analysis classifications from Blocks 2 and 3 to Blocks 4 and 5 of a single session. Jansen and van der Maas assume that consistent responding
within a single session implies abrupt transitions over months or years and that variable responding within a session implies gradual transitions. This assumption is plausible, but it is far from guaranteed.

To understand the uncertainty, consider the part of the idealized bottom panel of Jansen and van der Maas's Fig. 2 depiction that shows a child using Rule I for a substantial period of time and then switching abruptly to Rule II. The switch from Rule I to Rule II may be as abrupt as the depiction suggests. It also seems plausible, however, that if a given child's balance scale predictions were followed longitudinally over months or years, the switch from Rule I to Rule II would be found to occur gradually, with the child consistently using Rule I in one session and Rule II in a session a few months later but then switching back to Rule I in a yet later session. Such systematic regressions are not just a logical possibility. Siegler and Stern (1998) found that on arithmetic insight problems, 7- and 8-year-olds frequently discovered the insightful strategy and used it consistently for the remainder of the session, only to switch back to a slower and less accurate approach in the next session 1 week later. Thus, longitudinal data are needed to test Jansen and van der Maas's intriguing proposal regarding development on the balance scale task.

When Do Children Begin to Use Rules?

The age at which children can form grammatical rules has been the subject of a heated recent controversy (Marcus, Vijayan, Rao, & Vishton, 1999; Seidenberg & Elman, 1999). The age at which they can form problem-solving rules has received much less attention—until now. To address the issue, Halford, Andrews, Dalton, Boag, and Zielinski (2002, this issue) simplify the standard balance scale task in several ways that make it accessible to toddlers. In particular, they focus on problems on which only one of the two relevant balance scale dimensions (weight or distance) vary, and they present fewer items than were presented in previous rule assessment designs.

Halford et al.'s findings increase our understanding of initial acquisition of problem-solving rules in two main ways. One involves showing that, even before obtaining feedback experience with balance scales, 2-year-olds predict at above chance levels that the side with greater weight will go down. Siegler (1981) had shown that 3-year-olds who are presented with balance scale problems predict in accord with weight more often than by chance. These new results show that 2-year-olds also do so.

What conclusions should we draw from this finding? Halford et al. are almost certainly right that the 2- and 3-year-olds' knowledge is best viewed as a precursor to rule use rather than as rule use per se. The criterion that the 2-year-olds meet (above chance performance for the group as a whole) is much less stringent than the criterion of 80% of each child's responses generally used in rule assessment studies. Thus, Halford et al.'s findings tell us that the knowledge that weight influences the movement of balance scales has begun to develop by 2 years of age, but not that 2-year-olds understand the role of weight in the same systematic fashion as do 5-year-olds.
A second important conclusion that emerges from Halford et al.'s study is that, when presented with relevant experience, very young children can learn to consider dimensions that they ordinarily would not consider for years thereafter. In the absence of feedback and instruction regarding the relevant dimensions of the task, 5-year-olds do not even encode the role of distance, much less incorporate it into rules for predicting the balance scale's behavior (Siegler, 1976). Yet Halford et al.'s findings demonstrate that, given feedback and explanations of the role of distance, even 2-year-olds predict in accord with distance at above chance levels on problems where only distance varies. There is frequently a very large difference between what children do spontaneously and what they are capable of doing when given appropriate experience. Halford et al.'s findings illustrate just how large the difference can be.

*How Do Children Generate New Rules?*

Learning is a central part of children's lives. Indeed, childhood can be defined as the period of life during which learning is the central task. Much of the reason why we today view adolescence as part of childhood, rather than as part of adulthood, is that modern adolescents, unlike their peers of the past, have so much to learn before they can assume adult roles. Thus, although we usually think of childhood as a period defined by biology, it is also a social category defined in large part by the importance of learning relative to use of existing skills in self-sustaining activities.

Despite this central role in children's lives, learning has been relegated to a rather peripheral position in the recent history of research on cognitive development (Siegler, 2000). Consistent with this general tendency, studies of children's rule use have focused much more on what rules children use to solve problems than on how they acquire those rules.

The lead articles in this issue are consistent with this broader emphasis within the field. Jansen and van der Maas's study examines only children's spontaneous rule use. Halford et al. do include a learning phase, but they do not report any data from it. They describe whether children used weight and distance information on the pretest and on the posttest, but they do not describe how the children made the transition.

At least to us, the question of how the transition occurred in the Halford et al. study is at least as interesting as whether 2-year-olds can base responses on the balance scale task on the weight or distance dimension. Furthermore, relational complexity theory might be tested in a more precise way through analyses of the points in the learning process that produce difficulties for children of different ages than through global analyses of whether children learned or failed to learn. Examining the learning process also enhances our understanding of why 2-year-olds could learn to consider distance on problems on which only distance varied but not on problems on which both weight and distance varied.

We have begun to analyze the mechanisms that underlie acquisition of balance scale rules (Siegler & Chen, 1998). To do this, we have employed a microgenet-
ic design. This approach involves observing changing performance on a trial-by-trial basis as children encounter experience intended to promote learning. Microgenetic designs yield the type of high-density data needed to understand cognitive change (Siegler & Crowley, 1991).

In the Siegler and Chen (1998) study, 4- and 5-year-olds were presented with distance problems on which the two sides of the scale had the same amount of weight but on which the weight on one side was farther from the fulcrum. The goals were to see whether children acquired Rule II and, if so, how they did so. The study used a pretest–feedback–posttest design similar to that used by Halford et al., but performance during the feedback phase was subjected to intensive analysis designed to illuminate the learning process. On each of the 16 trials during the feedback phase, each child was asked to predict which side of the scale, if either, would go down if a lever that held the arm motionless were released. Following the child’s prediction, the lever was released and the child observed the scale’s movement. Finally, the child was asked to explain the outcome that occurred.

The trial-by-trial analysis of changes in children’s performance allowed us to identify four component processes that were involved in learning on this task and that are likely to be involved in learning on many other tasks as well. The first component was noticing potential explanatory variables (e.g., the role of distance) that had been ignored previously. To be classified as executing this component, children needed to refer to the location of the weights on the balance scale’s pegs on at least one occasion. The reference could be phrased either descriptively (“These weights are here, and these are here”) or predictively (“These weights are farther from the fulcrum, so this side will go down”). The second component was formulating a rule that incorporated distance as well as weight. To be classified as formulating a rule, children needed to explain the scale’s action on at least one trial by stating that a given side went down because its disks were farther from the fulcrum and then on the next trial predicting that the side with disks farther from the fulcrum would go down. The third component was generalizing the rule to novel problems by using it consistently after it was formulated. This was operationally defined as the child meeting the criteria for the formulating component on at least 80% of trials after the rule’s initial formulation. The fourth component was maintaining the new rule under less facilitative circumstances. This was operationally defined as using the new rule on at least 80% of trials on a no-feedback posttest the next day.

This componential analysis proved useful for understanding how children learn new balance scale rules. The key variable for learning in both age groups, and the largest source of developmental differences in learning, was noticing the potential explanatory role of distance from the fulcrum. Most 5-year-olds noticed the potential role of distance during learning, whereas most 4-year-olds did not. Children of both ages who noticed the role of distance were likely to formulate, generalize, and maintain the new rule; those of both ages who did not were not. The same componential analysis of children’s learning of scientific concepts has
proved useful in examining children's learning about water displacement (e.g., Siegler & Chen, in preparation) and seems applicable to acquisition of many other concepts as well. Thus, attention to how children acquire new rules can identify generalities beyond those indicated by their untutored performance.

CONCLUDING COMMENTS

The questions that we have raised, and the directions that we have recommended, in no way diminish the contributions of the lead articles. Halford and his colleagues (2002, this issue) demonstrate that the development of understanding of balance scales starts very early—by 2 years of age. They also show that when 2-year-olds are presented with feedback and verbal instructions regarding the role of distance from the fulcrum, they can learn that it too influences the balance scale's behavior. Jansen and van der Maas (2002, this issue) illustrate the usefulness of latent class analysis for examining developmental changes in rule use and find that systematic adherence to a single rule and variable strategy use both can occur within a single developmental progression. These findings extend the already substantial body of information we possess regarding children's knowledge about balance scales at different points in development. The challenge now is to increase our understanding of how children acquire this knowledge.

REFERENCES


Received July 19, 2001; revised August 18, 2001