Antti Lehtinen (Ph.D.) is a senior lecturer at the Department of Physics, University of Jyväskylä. His research focuses on physics education and teacher education, especially on physics instructional labs and teaching assistant training.

Salla Skyttä (MSc) is a doctoral researcher at the Department of Teacher Education, University of Jyväskylä. She graduated as a subject teacher with a major in mathematics. She teaches pedagogy of mathematics to primary school teachers. Her research interests include the flexible mathematical skills of primary school students and digital learning environments.

Pasi Nieminen (Ph.D.) is a senior lecturer at the Department of Teacher Education, University of Jyväskylä. His research areas include inquiry-based science learning and e.g. formative assessment, argumentation and differentiation.

Markus Hähkiöniemi (Ph.D., docent) is a senior lecturer at the Department of Teacher Education, University of Jyväskylä. His research interests include technology-enhanced inquiry-based mathematics teaching, argumentation, and classroom interaction.

ANTTI LEHTINEN
University of Jyväskylä, Finland
antti.t.lehtinen@jyu.fi

SALLA SKYTTÄ
University of Jyväskylä, Finland
salla.m.skytta@jyu.fi

PASI NIEMINEN
University of Jyväskylä, Finland
pasi.k.nieminen@jyu.fi

MARKUS HÄHKIÖNIEMI
University of Jyväskylä, Finland
markus.hahkioniemi@jyu.fi

Collaborative balance rule learning: Do students' age, group composition, prior knowledge, and scientific reasoning skills matter?

Abstract

Research on balance rule learning has focused on studies done in individual settings. This study investigates how students collaboratively learn balance rules and focuses especially on four variables that potentially affect rule development: student age, group composition, prior knowledge, and scientific reasoning skills. Eight-, ten- and twelve-year-old students collaboratively used a designed simulation-based learning environment with an open experimentation space and tasks that required progressively more *complex balance rules. Students' balance rules were tested before and after intervention with the Balance Scale Test and their scientific reasoning skills were tested with items from the Science-P Reasoning Inventory. The results show that the intervention was successful in developing students' balance rules. Logistic regression show that the students' previous knowledge was the only variable that affected the likelihood of rule development. Students' with less complex pre-test rules developed their rules more often than students with more complex pre-test rules when controlling for the other variables. The results go against some previous findings and show that a collaborative setting can lead to balance rule learning with primary school aged students.*

INTRODUCTION

This study investigates how students learn balance rules collaboratively, focusing on four variables that may influence rule development: student age, group composition, prior knowledge, and scientific reasoning skills. Balance rules are different mental procedures that people follow to balance a balance beam (Jansen & van der Maas, 2002). An example of a balance rule is ”The beam always turns to the side where the weight is farthest from the fulcrum”. In general, older children working individually use more sophisticated balance rules than younger children (Tourniere & Pulos, 1985; Siegler & Chen, 2002), but whether the same effect holds for collaborative learning scenarios remains to be investigated. In terms of group composition, some studies have found that homogeneous groups, i.e. groups in which participants have similar levels of prior knowledge, outperform heterogeneous groups (e.g. Fuchs et al., 1998; Hooper, 1992; Jensen & Lawson, 2011), while others have found evidence that heterogeneous groups outperform homogeneous groups (e.g. Saner et al., 1994; Pine & Messer, 1998; Webb et al., 1998, Webb et al., 2002). The positive effect of prior knowledge on learning is well documented (e.g. Dochy et al., 1999). Studies have also shown that scientific reasoning skills are central to rule learning (Siegler & Chen, 2002; Osterhaus et al., 2020).

Previous research on the development of balance rules has focused on students learning about balance with highly controlled study designs (e.g., Roth, 1991; van der Graaf, 2020). An open question is what balance rule development looks like in a more naturalistic setting, where students with different levels of knowledge work together to learn about balance. The results have implications for other areas of science education where different rules or explanatory models are considered, such as learning about DC circuits (Kokkonen & Mäntylä, 2018).

THEORETICAL BACKGROUND

Rule usage and balance

The scientific phenomenon of balance is often used to study people's use of rules (Hardiman et al., 1986; Normandeau et al., 1989; Siegler, 1976; Siegler & Chen, 2002). Rules are mental procedures that people follow to solve problems (Jansen & van der Maas, 2002). Research on children's use of rules and the balance beam can be traced back to the seminal work of Inhelder and Piaget (1958). Building on this, Siegler (1976, 1981, 1986) developed the Rule Assessment Methodology to build on the Piagetian method of cognitive assessment.

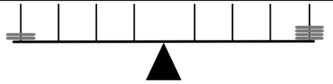
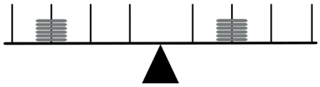
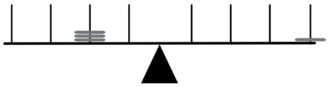
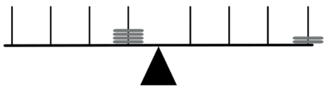
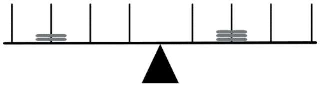
The Rule Assessment Methodology is based on two assumptions, the first being that children's reasoning is developmental and rule-governed. This assumption was later challenged by the so-called 'overlapping waves model', in which children were seen as often having multiple competing rules at their disposal, between which they may switch depending on contextual factors (Jansen & van der Maas, 2004; Siegler, 1996). Siegler (1976) outlined four rules: First, children compare only the weights on either side of the fulcrum (*Rule I*); second, they compare the distances at which the weights are placed from the fulcrum, but only when the weights on either side are equal (*Rule II*); third, they consider both dimensions but do not know how to combine them, so they guess or muddle through (*Rule III*); fourth, they learn to multiply the two dimensions and compare the products of both sides (*Rule IV*).

Since Siegler formulated these rules, alternative rule formulations have been proposed. Among these, Normandeau et al (1989) observed the addition rule and the qualitative proportionality (QP) rule. Children using the addition rule added the weight and the distance of the weight from the fulcrum from each side and compared the results. Children using the QP rule considered both weight and distance and concluded that a heavy weight at a short distance on one side of the fulcrum should compensate for a light weight at a greater distance on the other side of the fulcrum. They therefore predicted that the beam would remain horizontal in these cases.

Empirical evidence shows that older children use more sophisticated rules than younger children (Siegler & Chen, 2002; Tourniere & Pulos, 1985). Three-year-olds do not yet use rules often enough to solve balance-beam problems, whereas most five-year-olds do (Siegler, 1976, 1981). Most five to eight year olds use Rule I, while others use Rule II (Jansen & van der Maas, 2002). Nine to 12 year olds use either Rule I, Rule II, Rule III or the addition rule. It is not until the age of 13 or 14 that a significant number of children begin to use Rule IV. Furthermore, only a minority of adults use Rule IV, with most using either Rule III or the addition rule (Siegler & Chen, 2002).

Siegler's (1976, 1981) second assumption was that children's rules can be represented by a particular pattern of successes and failures in a series of problems. Rule use is then implied as an individual's consistent application of a particular rule to particular item types. Small deviations are allowed; typically, at least 80% of an individual's responses should follow a rule in order to be classified as such. Assessments of children's rule use are commonly based on five item types (Siegler, 1976). Table 1 shows these five item types and the expected proportion of correct responses for the six different balance rules.

Table 1 The expected portion of correct answers per item type for the six different balance rules

		Rule					
Item type	Example	Rule I	Rule II	Rule III	Rule IV	Addition rule	QP rule
Weight		1.00	1.00	1.00	1.00	1.00	1.00
Distance		.00[a]	1.00	1.00	1.00	1.00	1.00
Conflict-weight		1.00	1.00	.33[c]	1.00	.00[a]	.00[a]
Conflict-distance		.00[b]	.00[b]	.33[c]	1.00	1.00	.00[a]
Conflict-balance		.00[b]	.00[b]	.33[c]	1.00	1.00	1.00

Note. On this test, the addition rule results in the response "balance" to all conflict–weight items and in the correct response to all other conflict items

[a] Answers that the beam will stay in balance

[b] Answers that the beam will tip to the side with weights

[c] Guesses

Balance rule learning

The learning of balance rules has received less research attention than the use of balance rules (Siegler, 2000; Siegler & Chen, 2002). Different types of interventions have been used to study how learners learn new rules. One type of intervention involves learners either observing someone else manipulating a balance beam or manipulating it themselves without additional instruction (e.g., Hardiman et al., 1986; Li et al., 2017). Another type of intervention involves some form of instruction, such as verbal feedback or various balance tasks (e.g., Chletsos & De Lisi, 1991; Kliman, 1987; Li et al., 2017; Philips & Tolmie, 2007; Siegler & Chen, 1998; van der Graaf, 2020). In this section, we focus on previous research related to four variables: student age, group membership type, prior knowledge, and scientific reasoning skills.

Research has generally shown that older students benefit more from balance rule learning interventions than younger students. 14-year-old students were better able to maintain rule IV than 11-year-old students three months after a rule learning intervention (Chletsos & de Lisi, 1991). Eight-year-olds were more likely to form Rule III than five-year-olds when both groups were given feedback on balance problems and the complexity of the pre-test rule was controlled (Siegler, 1976). In another intervention study, four-year-olds did not reach Rule II as often as five-year-olds when the complexity of their pretest rule was controlled (Siegler & Chen, 1998). The reason for these results may lie in the development of general cognitive resources to process and store data and, in particular, the development of scientific reasoning skills (Koerber et al., 2005; Zimmerman, 2007). However, van der Graaf (2020) found a contradictory result with eight to 13 year olds. In their inquiry-based intervention, without controlling for prior knowledge, the younger learners learned more, i.e. they had a greater difference between the Balance Scale Test pre- and post-test scores.

With collaborative learning, the type of group membership of the students should also be considered. By collaborative learning, we refer to situations in which two or more learners are engaged in a common task or problem and use each other's resources and skills to solve it (Dillenbourg, 1999; van Aalst, 2013). By group membership type, we refer to the level of prior knowledge of an individual learner in relation to others in the same group. Some researchers have found that homogeneous groups (where all members of the group have similar amounts/types of prior knowledge on the topic) are more beneficial than heterogeneous groups (e.g., Fuchs et al., 1998; Hooper, 1992; Jensen & Lawson, 2011). One theoretical argument put forward to support these findings is based on Piaget's (1985) theory of equilibration, which posits that encountering new experiences during learning reorganises prior mental structures (e.g., rules) as gaps and inconsistencies are discovered through equilibration. This individual process is most effective when there is no interference or guidance from more capable peers. Other studies have found that heterogeneous groups outperform homogeneous groups (e.g. Pine & Messer, 1998; Saner et al., 1994; Webb et al., 1998, Webb et al., 2002). The theoretical rationale for these findings is based on Vygotsky's (1978) concept of the zone of proximal development, which posits that students perform better on tasks when they are grouped with more knowledgeable peers. The important point about the findings that heterogeneous groups outperform homogeneous groups is that the results are only valid for students who have a more knowledgeable peer in their group. Such findings suggest that students with higher prior knowledge benefit more from homogeneous groups (Webb et al., 1998; Webb et al., 2002). However, the only study dealing with balance and collaborative rule learning is Pine and Messer's (1998) study of five to seven year olds working in groups of four to solve balance scale problems. In the study, two types of groups were formed: heterogeneous groups with children demonstrating at least three levels of rule explicitness in relation to balance rules, and homogeneous groups with all children at the same level of rule explicitness. Rule explicitness refers to mainly the students' ability to balance beams with more complex combinations of weights and their distances (e.g., different weights on the other sides of the fulcrum) but also the students' ability to verbalize their thinking (Karmiloff-Smith, 1992). The groups received two types of intervention: They were either forbidden to discuss their reasoning with others, or they were encouraged to do just that. When the learners were encouraged to discuss their reasoning, those in the heterogeneous groups developed the level of explicitness of their balance rule more often than those in the homogeneous groups. Without discussion, there was no difference in learning between heterogenous and homogenous groups. These results highlight the role of discourse as an enabler of learning in heterogeneous groups.

Prior knowledge has generally a positive effect on learning (e.g. Dochy et al., 1999). However, these results are inconclusive in relation to the learning of equilibrium rules. In a study by Chletsos and de Lisi (1991) with 11- and 14-year-olds, there were no differences in the number of increasingly directive prompts or the number of manipulations of the apparatus required to articulate Rule IV between students with less and more complex pretest rules. Conversely, four- and five-year-olds who knew Rule I prior to an intervention acquired Rule II more often than children who did not know Rule I prior to an intervention (Siegler & Chen, 1998). Furthermore, Siegler (1976) found that children who used Rule I before the intervention were more likely to develop their rule when presented with problems that required the use of Rule II than when presented with problems that required the use of Rule III. The last two findings could be explained by the developmental sequence of rule learning: Children develop their rules in series, and their knowledge of a previous rule helps them to acquire the next, more complex rule.

The process of rule learning depends on students' ability to notice potential explanatory variables and their connections, and to formulate rules based on evidence (Siegler & Chen, 1998). This is often preceded by the generation of hypotheses based on either prior knowledge or the results of previous attempts (Klahr & Dunbar, 1988). The ability to understand the relationship between hypotheses or theory and evidence is the underlying skill of scientific reasoning (Kuhn, 2002; Osterhaus et al., 2020). Different components of scientific reasoning, such as students' ability to experiment, interpret evidence and understand patterns in evidence (Osterhaus et al., 2020), are central to the process of rule learning. Scientific reasoning skills develop during childhood (Koerber et al., 2005; Zimmerman, 2007), for example the ability to investigate the relationship between two variables can develop by the age of 10 (Kanari & Millar, 2004), but with appropriate guidance even seven-year-olds can design valid experiments (Chen & Klahr, 1999).

The present study

The study presented in this paper focuses on an intervention administered to students from three age groups (8-, 10-, and 12-year-olds) where they collaboratively used a simulation-based learning environment to develop their balance rules. Previous rule learning interventions have focused on individual settings (Chletsos & De Lisi, 1991; Hardiman et al., 1986; Kliman, 1987; Li et al., 2017; Philips & Tolmie, 2007; Siegler, 2000; Siegler & Chen, 1998; Siegler & Chen, 2002; van der Graaf, 2020). We focused on four variables and their possible effect on rule development: student age, group membership type, prior knowledge and scientific reasoning skills Based on the literature, we hypothesize that older students (Chletsos & de Lisi, 1991; Siegler, 1976; Siegler & Chen, 1998) and those with more developed scientific reasoning skills (Koerber et al., 2005; Zimmerman, 2007) are more often capable of developing their rules. Regarding the effect of prior knowledge (Chletsos & de Lisi, 1991; Siegler & Chen, 1998) and group membership type (Fuchs et al., 1998; Hooper, 1992; Pine & Messer, 1998; Webb et al., 2002), the literature is conflicted, so we were unable to formulate hypotheses.

The research questions are as follows:

1. How do the rules used by primary school students to balance a balance beam develop after participating in the collaborative intervention?
2. What is the effect of student age, group membership type, prior knowledge and scientific reasoning skills on balance rule development?

MATERIALS AND METHODS

The participants and study context

The data were collected from 12 primary school classes, four from each grade level (2nd, 4th, and 6th grade). The classes were situated in three primary schools (School 1: three second grade classes; School 2: three fourth grade classes, and School 3: one second grade class, one fourth grade class, and four sixth grade classes), all of which were in various suburbs of a middle-sized INSERT COUNTRY NAME HERE city. Participation in the study was voluntary for the students, and informed consent was obtained from their guardians. The complete data set, that is, the pre-test, intervention, and post-test, was collected from 147 students: 51 second graders ($M_{\text{age, 2nd grade}}$ = 8.45 years, $SD_{\text{age, 2nd grade}}$ = .32 years), 45 fourth graders ($M_{\text{age, 4th grade}}$ = 10.35 years, $SD_{\text{age, 4th grade}}$ = .30 years), and 51 sixth graders ($M_{\text{age, 6th grade}}$ = 12.45 years, $SD_{\text{age, 6th grade}}$ = .28 years).

The intervention

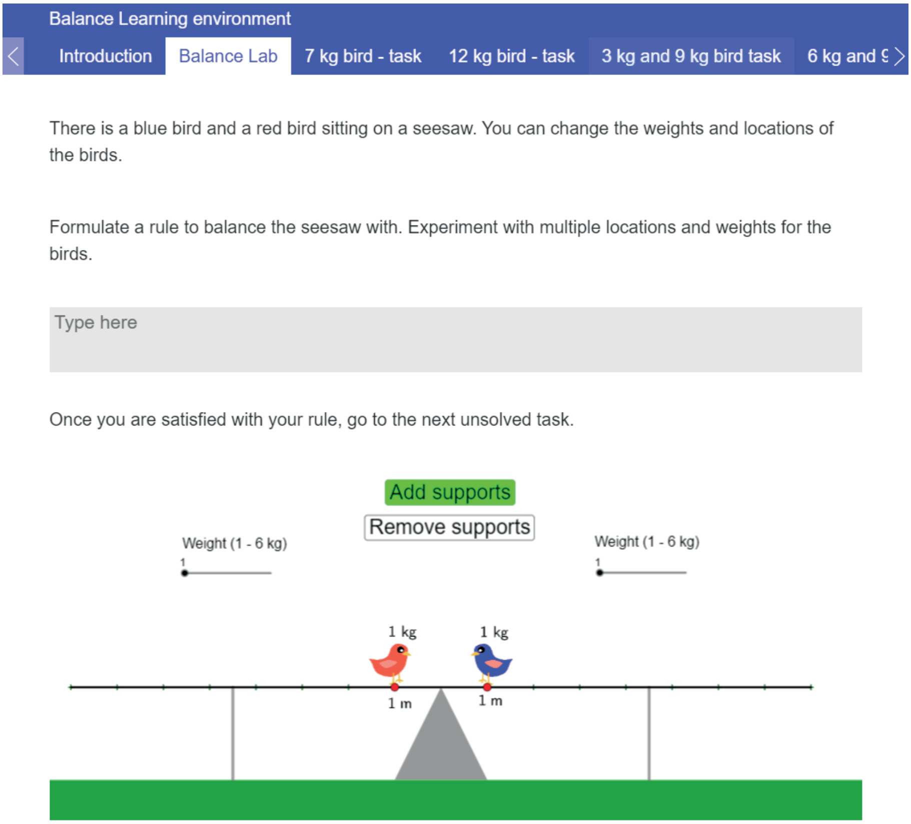
The data were collected from 63 small groups, 21 from each grade. For the intervention, the students from each class were randomly selected into groups of three. Due to the number of students in the class with research permits, the groups were not always divisible by three; therefore, 13 groups had two students. Each group was provided with a laptop, which ran the simulation-based learning environment, an external mouse, and a piece of paper with images of similar balance beams as those in the learning environment (for bookkeeping).

Even though the learning environment contained all the tasks and information needed to navigate the environments, guidance was provided due to the young age of the learners and the possibility that the transparency, that is, the ease of perceiving the content of the learning environment (Swaak et al., 1998) would be too low. Thus, one pre-service primary teacher (PST) was assigned to work with each group. Each PST worked with one group of second, fourth and sixth graders. Participation in the study was voluntary for the PSTs. The PSTs took part in the intervention as part of their science and mathematics education methods course where the focus was responsiveness to students' actions and adaptive support for learning. This meant that before the intervention all of the PSTs had practiced understanding students' ideas and providing them with guidance adapted to their thinking. For the intervention, the PSTs were instructed to focus on understanding the students' learning process and guide them based on their ideas. They were told not to provide the students with rules or strategies that the students had not yet verbalized, that is, to not provide new rules or strategies to the students.

The learning environment

A simulation-based learning environment was developed specifically for this study. The learning environment was piloted by testing two configurations (Author(s), 2022). In the learning environment, the students collaborated to construct a rule that could be used to balance a balance beam. Using simulations, the students could experiment with a balance beam where two birds of varying weight could be placed on different sides of the fulcrum and at different distances from the fulcrum. The learning environment was built with Graasp (2022), and the simulations were designed using GeoGebra (2022). The same learning environment was used with all age groups. This was taken into account when designing the learning environment by having the learning environment to require only simple balance rules at first and by using only small whole numbers for the weights and distances of the balance beam.

The learning environment consisted of seven tabs, the first of which instructed the students how to manipulate the simulation. The second tab was the Balance Lab (Figure 1), where the students could experiment with a balance beam. They were prompted to formulate a rule for balancing the beam based on their experiments to a text box.



There is a blue bird and a red bird sitting on a seesaw. You can change the weights and locations of the birds.

Formulate a rule to balance the seesaw with. Experiment with multiple locations and weights for the birds.

Type here

Once you are satisfied with your rule, go to the next unsolved task.

Figure 1 The balance lab tab in the learning environment

The four remaining tabs contained Tasks 1–4, which were aimed at testing the rule(s) the students had formulated. In each task, the students were presented with the rule(s) they had formulated in the Balance Lab tab and a simulation-based task where they had to balance a balance beam containing some fixed weights or distances. If the students succeeded in balancing the seesaw, they were instructed via text to move on to the next tab. If they were unsuccessful, they were instructed via text to return to the Balance Lab to try to formulate another rule. The weights in the tasks differed from those in the Balance Lab, so the students could not replicate the task situation in the Balance Lab. The final tab contained additional tasks in the "Balancing Act" PhET simulation (PhET Interactive Simulations, 2022). The purpose of these tasks was to provide the students with extra activities if they completed all the other tasks in the time allotted.

The learning environment was designed to balance the structuring and problematization of learning (Reiser, 2004). The principle of structuring learning was apparent in the tasks that required progressively more complex rules (see Table 2). For example, Task 1 could be solved by simply matching the weights and distances on both sides, that is, Rule II, but the later tasks required the students to consider both the weights and distances by increasing the complexity of the ratios. We have named the principle of combining the mostly unguided experimentation environment (i.e., the Balance Lab) and guidance provided through tasks that require increasingly more complex rules to solve as implicit model progression (Author(s), 2022). The principle of problematizing learning was apparent in the fact that if the students failed a task, they were prompted to return to the Balance Lab and further develop their rule instead of simply bypassing the task.

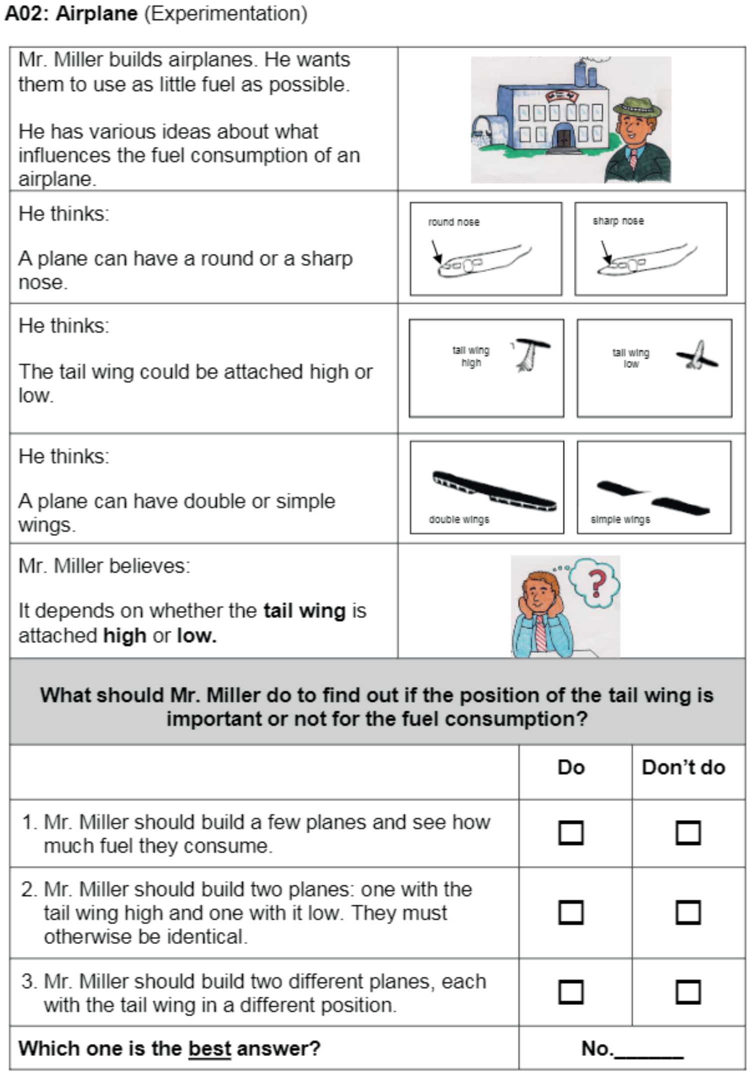
Table 2 The weights and distances of the birds in the Balance lab and tasks

	Bird on the left		Bird on the right	
	weight	distance	Weight	distance
Balance Lab	1–6 kg	1–8 m	1–6 kg	1–8 m
Task 1	7 kg	2 m	1–20 kg	1–8 m
Task 2	12 kg	1 m	1–7 kg	1–8 m
Task 3	3 kg	2–8 m	9 kg	2–8 m
Task 4	6 kg	1–8 m	9 kg	4 m

Data collection

The students' rules for balancing the balance beam were assessed before and after the intervention using the well-documented Balance Scale Task (BST) (Jansen & van der Maas, 2002; van Maanen et al., 1989). The BST has good internal consistency, and it is usable even with learners as young as five-years old (Jansen & van der Maas, 2002). The BST consists of five blocks of five items each, and the items in each block are each of a different problem type. The items were arranged in the same order in each block: weight, distance, conflict–weight, conflict–distance, and conflict–balance. The conflict items were designed in such a way that the use of the addition rule would result in a correct response to the conflict–distance and conflict–balance items but an incorrect response of "in balance" to the conflict–weight items.

The students' scientific reasoning skills were assessed using items from the reduced Science-P Reasoning Inventory (SPR-I(7)) (Osterhaus et al., 2020). The SPR-I and SPR-I(7) were designed to assess primary-school students' scientific reasoning skills across three components: experimentation, data interpretation, and understanding the nature of science. The SPR-I and SPR-I (7) can be found at: https://osf.io/34dsk/. Due to a) the importance of experimentation and data interpretation in rule learning and b) time constraints during data collection, we only used the four items related to experimentation and data interpretation from SPR-I(7). The three experimentation items addressed the control-of-variables strategy (see Figure 2 for an example item) and the differentiation between the production of an effect and the test of a hypothesis. The single data interpretation item addressed the understanding of confounded data patterns. All items had three answer options corresponding with three levels of understanding: naïve, intermediate, and advanced. For each option, the students were asked to indicate their agreement or lack thereof. Items were scored as zero points when the student selected the naïve answer or when they rejected all answer options. Students who selected the intermediate but not the naïve answer were given one point and those that selected the advanced answer but rejected all other answers were given two full points. The sum of the score from all items (zero to eight) was used to measure the students' scientific reasoning skills. The mean score was 3.52 (SD = 1.79). The mean score for the second graders was 2.70 (SD = 1.59), for the fourth graders 4.07 (SD = 1.76), and for the sixth graders 3.85 (SD = 1.76).



A02: Airplane (Experimentation)

Mr. Miller builds airplanes. He wants them to use as little fuel as possible. He has various ideas about what influences the fuel consumption of an airplane.	
He thinks: A plane can have a round or a sharp nose.	round nose sharp nose
He thinks: The tail wing could be attached high or low.	tail wing high tail wing low
He thinks: A plane can have double or simple wings.	double wings simple wings
Mr. Miller believes: It depends on whether the **tail wing** is attached **high** or **low.**	

What should Mr. Miller do to find out if the position of the tail wing is important or not for the fuel consumption?

	Do	**Don't do**
1. Mr. Miller should build a few planes and see how much fuel they consume.	□	□
2. Mr. Miller should build two planes: one with the tail wing high and one with it low. They must otherwise be identical.	□	□
3. Mr. Miller should build two different planes, each with the tail wing in a different position.	□	□
Which one is the best answer?	**No.______**	

Figure 2 One of the experimentation items from the SPR-I (7) (Osterhaus et al., 2020)

Figure 3 showcases the course of the data collection and the intervention. The intervention lessons lasted about 50 minutes. BST data was collected before and after the intervention and the SPR-I (7) data was collected after the intervention. As scientific reasoning skills are expected be quite stable, we expect that the intervention itself did not have an effect on the students' scientific reasoning skills.

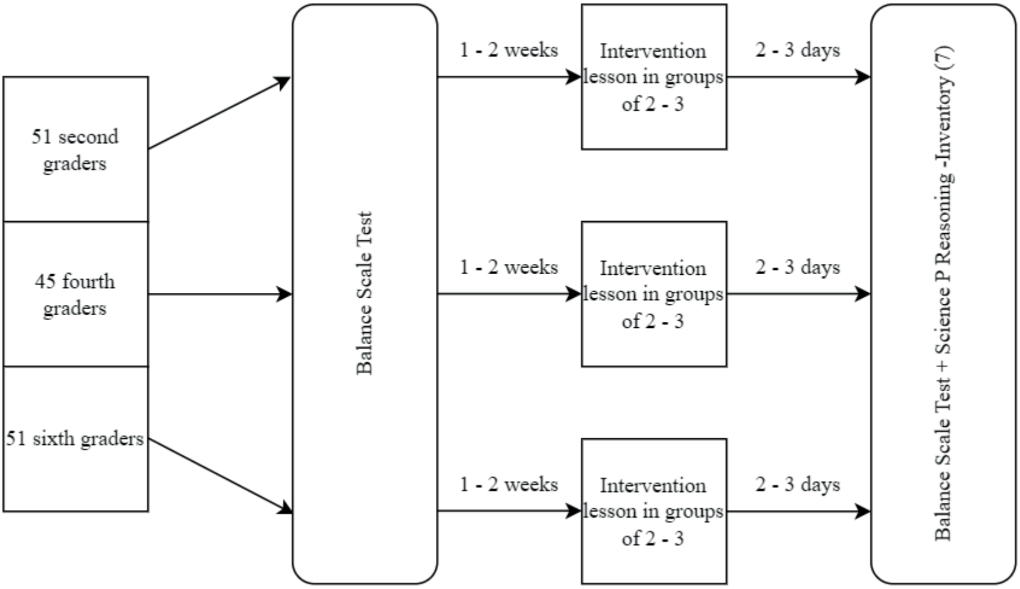

51 second graders
45 fourth graders
51 sixth graders
Balance Scale Test
1 - 2 weeks
Intervention lesson in groups of 2 - 3
2 - 3 days
Balance Scale Test + Science P Reasoning -Inventory (7)

Figure 3 The course of the data collection and the intervention

Analysis of the Balance Scale Test -data

Jansen and van der Maas (2002) found that children's answers for the first five problems in the BST differed from their answers in the remaining problems, which may be due to them encountering these sorts of problems for the first time. Consequently, Jansen and van der Maas discarded the first five problems from their data. For this study, we initially used the rule assessment methodology with both the full set of 25 items and then the latter 20 items (i.e., four items per problem type). By using only 20 items, we were able to assess more students who used one of the six rules. The results and a similar experience by Jansen and van der Maas (2002) led us to use only the results from the latter 20 items.

We analyzed the data for rules I–IV (Siegler, 1976), the addition rule, and the QP rule (Normandeau et al., 1989). The original formulations for the combinations of answers to various problem types were scaled to correspond with the BST, with 20 items in total and four items per problem type. To be assessed as having used Rule I meant that each student had to answer to at least 17 problems according to Rule I as well as answer "balance" to at least three of the distance problems. For Rule II, the student had to answer at least 17 problems according to Rule II, and from these answers, at least three had to be to the distance problems. For Rule III, the student had to answer correctly to at least seven non-conflict problems, of which at least three had to be distance problems. Furthermore, the student had to answer correctly to a maximum of nine conflict problems. To be assessed as having used the addition rule, the student had to answer at least 17 problems according to the addition rule, of which at least three answers had to be according to the addition rule from each problem type. For the QP rule, the student had to answer at least 17 problems according to the QP rule, of which at least three answers had to be according to the QP rule from each problem type. For Rule IV, the student had to answer correctly to at least 17 problems. The BST answers by some students did not correspond to any of the six rules. These students were assessed as having used an *unclassified rule or no rule at all*. Similar results were found in Jansen and van der Maas (2002) and van der Graaf (2020).

Similar to Normandeau et al. (1989), some of the children in this study could be assessed as having used Rule III and either the QP or addition rule. In our sample, all those using either the QP or addition rule could be assessed as having also used Rule III. The converse relation was not true: Some students were assessed as only having used Rule III. In these cases, we followed Normandeau et al. (1989) and ordered the different rules based on their complexity: (unclassified rule or no rule) < Rule I < Rule II < Rule III < Addition rule < QP rule < Rule IV. In cases where a student was assessed to have used two rules, the rule reflecting the highest level of ability was used in further analyses.

Three group membership types were used in the analysis: 1) the student was part of a heterogenous group, with their pre-test rule being the most complex (or tied with another learner's pre-test rule) in the group; 2) the student was part of a heterogenous group, and their pre-test rule was not the most complex in the group; and 3) the student was part of a homogenous group in which all members had the same pre-test rule.

Analysis of balance rule development

To answer research question 1, the Wilcoxon signed-rank test was run to study the statistical significance between the rules assessed from the pre- and post-tests. Wilcoxon signed-rank test is suitable for as the rules were assessed on an ordinal level. To answer research question 2, rule development was conceptualized as a change from a less complex rule in the pre-test to a more complex one in the post-test. This meant that students were categorized into two groups: those who developed their rule during the interventions and those who did not. Thus, a logistic regression analysis was used. Logistic regression estimates the probability of an event occurring (e.g., a student's rule developed during the intervention, or it did not develop), based on a given dataset of independent variables. Our independent variables were student age, prior knowledge, scientific reasoning skills and group membership type. For the sake of brevity, student age was operationalized as grade level. Prior knowledge was operationalized as the students' pre-test rule.

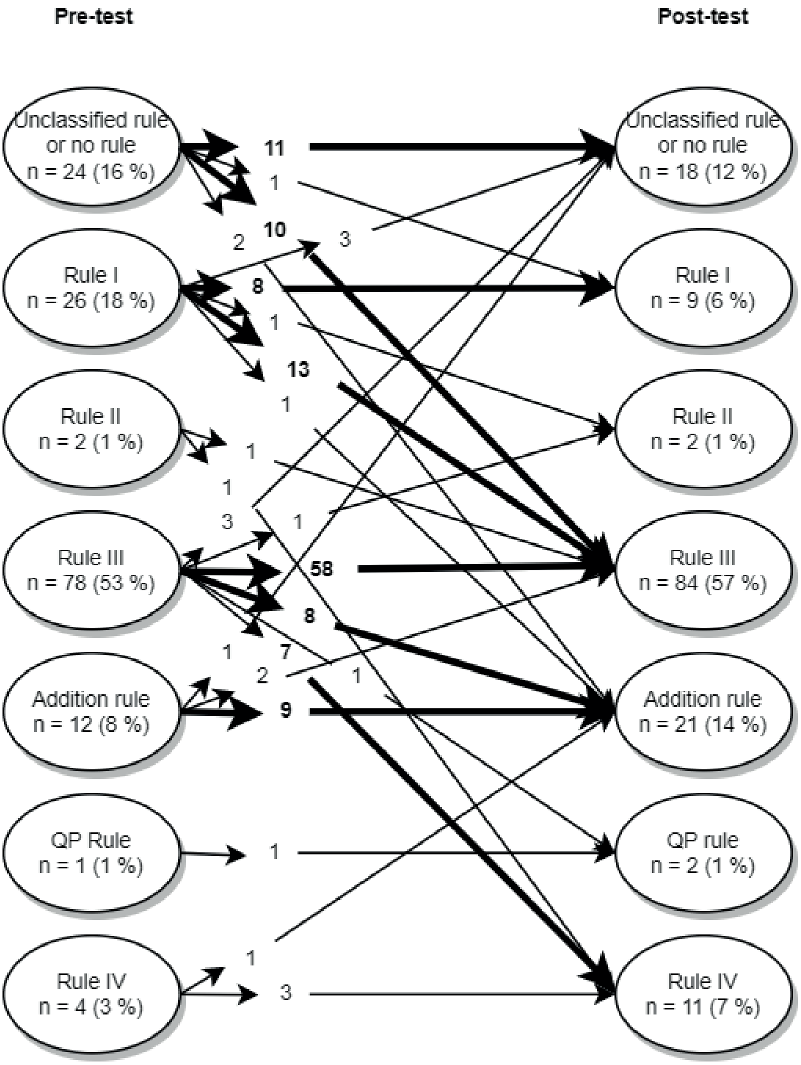
Students who were assigned Rule IV in the pre-test (n = 4) were excluded from the analyses because rule development was impossible for them. Seven students in four groups from which one student was assigned an unclassified rule or no rule at all in the pre-test and other students were assigned Rule I were also excluded from the analysis. The rationale for this exclusion criteria was that the students who were assigned an unclassified rule or no rule often had a higher percentage of correct items in the BST than those who were assigned Rule I in the pre-test. Similar exclusion principles have been used in the rule learning literature (Chletsos & de Lisi, 1991). After these exclusions, data from 128 students were used to answer research question 2.

RESULTS

The balance rules before and after the intervention

Figure 4 displays how many students were assessed as having used each rule in the pre-test (left side of the figure) and post-test (right side of the figure). Before the intervention, most students were using Rule III or Rule I, an unclassified rule, or no rule at all. After the intervention, the most common rule was again Rule III, but the addition rule and an unclassified or no rule were also prevalent. 31 % of the students used a more complex rule, 61% used the same rule and 7 % used a less complex rule after the intervention.

The distribution of rules assessed from the post-test was different than what was assessed from the pre-test and the Wilcoxon signed-rank test showed that the difference was statistically significant ($Z = -4.526$, $p < .001$). The distribution of rules assessed from the post-test was more inclined more towards complex rules.



Note: The bolded transitions contained more than 5% of the students.

Figure 4 The rules from the pre- and post-tests and the transitions between them for all students (n = 147)

The effect of student age, prior knowledge, scientific reasoning skills, and group membership type on balance rule development

For the logistic regression, the students' pre-test rules were divided into two categories, The pre-test rule was either less complex than Rule III, or it was Rule III or more complex. This was done to have sufficiently large groups with which to run the analysis. The cutting point (i.e., Rule III) was chosen because it was the simplest rule where the students considered both dimensions. Table 3 shows the descriptive data on the percentage of students who developed their rule for different values of the categorical variables.

Table 3 The percentage of students who developed their rules divided by the values of the categorical variables

		n	Percentage of students who developed their rule
Grade	**2nd**	**40**	**43%**
	4th	**41**	**24%**
	6th	**47**	**30%**
Pre-test rule	**Less complex than Rule III**	**41**	**61%**
	Rule III, Addition rule or QP rule	**87**	**19%**
Group membership type	**Most complex pre-rule**	**48**	**15%**
	Not most complex pre-rule	**50**	**46%**
	Homogenous	**30**	**36%**

The log-likelihood ratio test was used to compare the full logistic regression model (which contained four variables) with models containing either three or two variables. A comparison between the full model with all four variables and the models containing three variables revealed that the full model fit the data better than the models without scientific reasoning skills (χ^2 (1) = 10.411, p < .002), without the pre-test rule (χ^2 (1) = 11.833, p < .001), and without group membership type (χ^2 (2) = 13.514, p < .002). However, a comparison between the full model and the model without the grade variable revealed no statistically significant difference (χ^2 (2) = 1.642, p = .44). Thus, the grade variable was dropped from the model. The model containing three variables was then compared with those containing two variables. The three-variable model fit the data better than the model without scientific reasoning skills (χ^2 (1) = 9.968, p < .002), without the pre-test rule (χ^2 (1) = 21.841, p < .001), and without group membership type (χ^2 (2) = 14.015, p < .001). Thus, the model containing three variables was chosen for the analysis.

Table 4 showcases the results from the logistic regression. The model was statistically significant, χ^2(4) = 27.666, p < .001, and explained 27.2% (Nagelkerke R^2) of the variance in rule development and correctly classified 74.2% of the cases. When controlling for the other variables, students with rules less complex than Rule III in the pre-test were 9.7 times more likely to develop their balance rules than those with Rule III, Addition rule or QP rule in the pre-test. There were no statistically significant differences between the likelihood of developing the rules between students from different group membership types. Furthermore, scientific reasoning skills did not affect the likelihood of the students developing their balance rules.

Table 4 The logistic regression model

	Predictor	β	SEβ	Wald's χ^2	df	p	Odds ratio
Pre-test rule	**Less complex than Rule III (v. Rule III, Addition rule or QP rule)**	**2.272**	**.653**	**12.099**	**1**	**<.001***	**9.698**
Group membership type	**Most complex pre-rule (v. homogenous)**	**-.939**	**.585**	**2.574**	**1**	**.109**	**.391**
	Not most complex pre-rule (v. homogenous)	**-.793**	**.658**	**1.456**	**1**	**.228**	**.452**
	Not most complex pre-rule (v. most complex)	**.145**	**.715**	**.041**	**1**	**.839**	**1.156**
Scientific reasoning		**.179**	**.128**	**1.961**	**1**	**.161**	**1.196**

* Significant, $p < .05$

DISCUSSION

The effect of the intervention on the balance rules

The results indicate that the collaborative intervention was successful in developing the students' balance rules with 31% of the students using a more complex rule after the intervention. Rule III was the most frequently used rule before and after the intervention. This result differs from that of Jansen and van der Maas (2002), where the use of Rule III was uncommon in the same age groups. Before the intervention, the use of Rule IV was scarce, which is in line with results from previous studies (Jansen & van der Maas, 2002; Siegler, 1976, 1981).

These results can be compared to those of van der Graaf (2020), with 8–13-year-olds individually participating in an inquiry-based intervention also using a simulation-based learning environment. 26 % of the students in van der Graaf's study developed their rule measured by BST. Rule IV was not present at all in the students' answers after the intervention. Van der Graaf even noted that the *"acquisition of the most complex strategy [Rule IV] in this age group [8- to 13-years-old] is unlikely after a single inquiry-based lesson."* (p. 11). Our results show that the acquisition of Rule IV is possible in that age group after a short collaborative intervention.

Individual students' transition from one rule to another due to the intervention can be used to assess how well the intervention succeeded in developing the balance rules for students with different pre-test rules. None of the students transitioned from using the addition rule to a more complex rule. This is an unexpected result because only Task 1 in the learning environment could be solved by the addition rule. It seems that the addition rule was generally resistant to change, corroborating the results of van der Graaf (2020). The use of Rules I and II diminished because of the intervention. This is a sign that the learning environment worked as expected because all the tasks required the consideration of both the weight and distance.

The effect of student age, group membership type, prior knowledge, and scientific reasoning skills on balance rule development

The results indicate that students with pre-test rules that were less complex than Rule III developed their rules more often than other students. This finding adds to the literature on the effect that the complexity of students' prior rules has on learning balance rules. Previous research on the effect of prior knowledge on rule learning has focused on its effect on the attainment of a particular rule, for example, Rule II (Siegler, 1976; Siegler & Chen, 1998) or Rule IV (Chletsos & de Lisi, 1991). We conceptualized rule learning as moving to use any more complex rule because of the intervention. As the learning environment was designed to support students in constructing increasingly complex rules through implicit model progression (Author(s), 2022), it provided opportunities for students with different pre-test rules to develop their rules. As a side effect, this design choice also meant that the number of tasks requiring students to use more complex rules than their pre-test rule was higher for those students with less complex pre-test rules. Students with Rule III or more complex pre-test rules were not immediately required to challenge and develop their existing rule by the learning environment. This could be avoided with a diagnostic assessment embedded in the learning environment or enacted by a teacher. Students with less complex prior rules could start immediately with the tasks that challenge their rule.

Student age was not a variable in the final logistic regression model. In previous research, older students have developed their balance rules more often (Chletsos & de Lisi, 1991), even when controlling for pretest rules (Siegler, 1976; Siegler & Chen, 1998). Our results can be seen contradicting both these findings and our original hypothesis. One possible reason for this that is partly supported by the results of this study is that moving from rule I to rule III is easier than moving from rule III to rule IV. As the simpler rules were more prevalent among the younger learners in the pre-test, this may have influenced the results. Another possible explanation is that the collaborative learning setting may have supported the younger learners in developing their rules. A larger dataset would allow one to study the effects of pre-test rule and student age for rule learning in more detail.

The results indicate that students' scientific reasoning skills had no effect on the likelihood that they would develop their balance rules. This finding contradicts our initial hypothesis. It may be that the collaborative design of the intervention prevented students with higher scientific reasoning skills from using the learning environment as they would have been able to via e.g., controlling for variables as now the group had to negotiate the use of the learning environment. Future research could, for example, investigate the role of the average scientific reasoning skills of each small group as a factor in rule development. The groups were also supported by the PSTs whose guidance may have been more directed towards the students with lower scientific reasoning skills.

The results of this research indicate that the type of group membership had no effect on the likelihood of students developing their balance rules. This finding can be viewed from three perspectives. First, when controlling for student age, scientific reasoning ability, and prior knowledge, students did not benefit from having a more able peer working with them. This result is somewhat surprising and contradicts some previous findings (Pine & Messer, 1998; Saner et al., 1994; Webb et al., 1998, Webb et al., 2002) and the notion of the zone of proximal development (Vygotsky, 1978). Second, in the data there was no difference in the likelihood of students developing their rules between members of heterogeneous and homogeneous groups, which also contradicts some previous findings (Fuchs et al., 1998; Hooper, 1992; Jensen & Lawson, 2011) and the theory of equilibrium (Piaget, 1985). All in all, these results support the notion that grouping students based on their previous ability does not have a significant effect on learning. Third, no difference appeared between members of the heterogeneous groups in the likelihood of rule development. This was a positive indicator from an equity perspective, as the literature has highlighted that high ability students' learning may be hindered in heterogeneous groups (Webb et al., 2002). Research has shown that the quality of group interaction is a stronger predictor of performance in collaborative learning than student ability or the ability composition of the group (Webb et al., 2002), and that outcomes related to group composition and learning are mediated by the amount and type of interaction between group members (Dillenbourg, 1999; Pine & Messer, 1998; Webb et al., 2002). For high-ability students working in heterogenous groups, high quality interaction within the group might promote learning to the same as it would be in homogenous groups (Webb et al., 2002). It is also possible, for example, that if higher quality interactions between students were supported by prompts or scripts, students with less complex pre-test rules might start to benefit from having a more capable peer in their group. It is also possible that the presence of the PST may have influenced the role of rule composition in the results.

Implications

The results of this research indicate that collaborative learning using a learning environment is beneficial for different-aged learners, including even eight-year-olds. In this context, grouping students to homogenous or heterogenous groups had no significant impact on rule development. Students with a more complex pre-test rule developed their rule less often but the design of the simulation-based learning environment might have affected this. More research with larger datasets is needed to validate this study's results regarding group composition and scientific reasoning skills. Regarding research, even though the current study focuses on balance rule learning, the results have implications for other areas of science education research. For example, there is a multitude of research on the explanation models that students use to explain the behavior of DC circuits (e.g., Kokkonen & Mäntylä, 2018; Koponen & Huttunen, 2013). The methodology applied in this study could also be expanded to study how students learn about electricity and whether that is affected only be previous knowledge as this study implies.

Limitations

One limitation of this study is the guidance provided by the PSTs. Even though the PSTs were instructed not to provide rules and all the students from all the grades were supported by the same 21 PSTs, individual groups might have received different forms of guidance. Another limitation is that due to the size of the data set, data-based methods could not be used to discern the rules used by the students, which could have uncovered additional rules used by them. A larger data set would have increased the statistical power of the analysis. One final limitation is that the pre-test data was not used to group students together but instead the groups were formed randomly.

ACKNOWLEDGMENTS

The authors want to thank the students and pre-service teachers for participating in the study. Funding for this study was provided by the Research Council of Finland (Grant number 318010).

DECLARATION OF INTEREST STATEMENT

We wish to confirm that there are no known conflicts of interest associated with this publication.

REFERENCES

Author(s). 2022

Chen, Z., & Klahr, D. (1999). All other things being equal: Acquisition and transfer of the control of variables strategy. *Child Development, 70*(5), 1098–1120.

Chletsos, P. N., & De Lisi, R. (1991). A microgenetic study of proportional reasoning using balance scale problems. *Journal of Applied Developmental Psychology, 12*(3), 307–330.

Dillenbourg, P. (1999). What do you mean by collaborative learning? In P. Dillenbourg (Ed.), *Collaborative-learning: Cognitive and Computational Approaches* (pp. 1–19). Elsevier.

Dochy, F., Segers, M., & Buehl, M. M. (1999). The relation between assessment practices and outcomes of studies: The case of research on prior knowledge. *Review of Educational Research, 69*(2), 145–186.

Fuchs, L. S., Fuchs, D., Hamlett, C. L., & Karns, K. (1998). High-achieving students' interactions and performance on complex mathematical tasks as a function of homogeneous and heterogeneous pairings. *American Educational Research Journal, 35*(2), 227–267.

GeoGebra Math Apps. (2022) http://geogebra.org

Graasp - A Space for Everything (2022) https://graasp.eu/

Hardiman, P. T., Pollatsek, A., & Well, A. D. (1986). Learning to understand the balance beam. *Cognition and Instruction, 3*(1), 63–86.

Hooper, S. (1992). Effects of peer interaction during computer-based mathematics instruction. *The Journal of Educational Research, 85*(3), 180–189.

Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence.* Basic Books. (Original work published 1955)

Jansen, B. R., & van der Maas, H. L. (2002). The development of children's rule use on the balance scale task. *Journal of Experimental Child Psychology, 81*(4), 383–416.

Jensen, J. L., & Lawson, A. (2011). Effects of collaborative group composition and inquiry instruction on reasoning gains and achievement in undergraduate biology. *CBE—Life Sciences Education, 10*(1), 64–73.

Kanari, Z., & Millar, R. (2004). Reasoning from data: How students collect and interpret data in science investigations. *Journal of Research in Science Teaching, 41*(7), 748–769.

Karmiloff-Smith, A. (1992). Beyond modularity: A developmental perspective on cognitive *science*. MIT Press.

Klahr, D., & Dunbar, K. (1988). Dual space search during scientific reasoning. *Cognitive Science, 12*(1), 1–48.

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

Koerber, S., Sodian, B., Thoermer, C., & Nett, U. (2005). Scientific reasoning in young children: Preschoolers' ability to evaluate covariation evidence. *Swiss Journal of Psychology, 64*(3), 141–152.

Kokkonen, T., & Mäntylä, T. (2018). Changes in university students' explanation models of DC circuits. *Research in Science Education*, 48, 753-775.

Koponen, I. T., & Huttunen, L. (2013). Concept development in learning physics: the case of electric current and voltage revisited. *Science & Education*, 22(9), 2227–2254.

Kuhn, D. (2002). What is scientific thinking and how does it develop? In U. Goswami (Ed.), *Blackwell handbook of childhood cognitive development* (pp. 371–393). Blackwell Publishers.

Li, F., Xie, L., Yang, X., & Cao, B. (2017). The effect of feedback and operational experience on children's rule learning. *Frontiers in Psychology, 8*, 1-8.

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

Osterhaus, C., Koerber, S., & Sodian, B. (2020). The Science-P Reasoning Inventory (SPR-I): Measuring emerging scientific-reasoning skills in primary school. *International Journal of Science Education, 42*(7), 1087–1107.

PhET Interactive Simulations. (2021). *Balancing Act*. https://phet.colorado.edu/en/simulation/balancing-act

Philips, S., & Tolmie, A. (2007). Children's performance on and understanding of the balance scale problem: The effects of parental support. *Infant and Child Development: An International Journal of Research and Practice, 16*(1), 95–117.

Piaget, J. (1985). The equilibration of cognitive structures: The central problem of intellectual development. University of Chicago Press.

Pine, K. J., & Messer, D. J. (1998). Group collaboration effects and the explicitness of children's knowledge. *Cognitive Development, 13*(1), 109–126.

Reiser, B. J. (2004). Scaffolding complex learning: The mechanisms of structuring and problematizing student work. *Journal of the Learning Sciences*, 13(3), 273-304.

Roth, W. M. (1991). The development of reasoning on the balance beam. *Journal of Research in Science Teaching, 28*(7), 631–645.

Saner, H., McCaffrey, D., Stecher, B., Klein, S., & Bell, R. (1994). The effects of working in pairs in science performance assessments. *Educational Assessment, 2*(4), 325–338.

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

Siegler, R. S. (1981). Developmental sequences within and between concepts. *Monographs of the Society for Research in Child Development, 6*(2).

Siegler, R. S. (1986). Unities across domains in children's strategy choices. In M. Perlmutter (Ed.), *Perspectives on intellectual development: Minnesota symposia on child psychology* (Vol. 19, pp. 1–8). Erlbaum.

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

Siegler, R. S. (2000). The rebirth of children's learning. *Child Development, 71*, 26–35.

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

Siegler, R. S., & Chen, Z. (2002). Development of rules and strategies: Balancing the old and the new. *Journal of Experimental Child Psychology, 81*(4), 446–457.

Swaak, J., van Joolingen, W. R., & de Jong, T. (1998). Supporting simulation-based learning: The effects of model progression and assignments on definitional and intuitive knowledge. *Learning and Instruction, 8*(3), 235–252.

Tourniere, F., & Pulos, S. (1985). Proportional reasoning: A review of the literature. *Educational Studies in Mathematics, 16*, 181–204.

Van Aalst, J. (2013). Assessment in collaborative learning. In C. E. Hmelo-Silver, C. A. Chin, C. K. K. Chan, & A. O'Donnell (Eds.), *The international handbook of collaborative learning* (pp. 280–296). Routledge.

Van der Graaf, J. (2020). Inquiry-based learning and conceptual change in balance beam understanding. *Frontiers in Psychology*, *11*.

Van Maanen, L., Been, P., & Sijtsma, K. (1989). The linear logistic test model and heterogeneity of cognitive strategies. In E. E. Roskam (Ed.), *Mathematical psychology in progress* (pp. 267–288). Springer-Verlag.

Vygotsky, L. (1978). Mind in society: Development of higher psychological processes. Harvard University Press.

Webb, N. M., Nemer, K. M., Chizhik, A. W., & Sugrue, B. (1998). Equity issues in collaborative group assessment: Group composition and performance. *American Educational Research Journal, 35*(4), 607–651.

Webb, N. M., Nemer, K. M., & Zuniga, S. (2002). Short circuits or superconductors? Effects of group composition on high-achieving students' science assessment performance. *American Educational Research Journal, 39*, 943–989.

Zimmerman, C. (2007). The development of scientific thinking skills in elementary and middle school. *Developmental Review, 27*(2), 172–223.