Instr Sci (2015) 43:381–400
DOI 10.1007/s11251-015-9344-y

Scientific reasoning abilities in kindergarten: dynamic assessment of the control of variables strategy

Joep van der Graaf · Eliane Segers · Ludo Verhoeven

Received: 2 June 2014 / Accepted: 12 January 2015 / Published online: 24 January 2015

Abstract A dynamic assessment tool was developed and validated using Mokken scale analysis to assess the extent to which kindergartners are able to construct unconfounded experiments, an essential part of scientific reasoning. Scientific reasoning is one of the learning processes happening within science education. A commonly used, hands-on, experimentation task was adapted to dynamically assess the use of the so-called control of variables strategy (CVS) by children 4 to 6 years of age. In this task, the children were challenged to design experiments using two ramps with up to four independent variables: weight of the ball, steepness of the slope, place of the starting gate, and surface texture of the slope. There were two scores of CVS use: experiment and variable correct score. The analysis showed it was possible to assess CVS use in a reliable and valid manner with the new assessment tool. Irrespective of the number of variables children were allowed to set, experiments validly measured CVS use. Given that the number of variables to be set increased the difficulty of the experiment, this can be used to scale children's CVS use. In other words, it is possible to differentiate between children on the basis of their CVS use. The children's use of CVS positively related to both their age and nonverbal reasoning ability. The present results thus show that it is feasible to evaluate the ability of kindergartners to construct unconfounded experiments using dynamic assessment. This means that kindergartners can use the CVS and might be seen as natural scientists, who can and will try to unravel the physical world around them. Their explorations appear to need sufficient guidance (i.e. within their zone of proximal development) to design multivariable experiments.

Keywords Control of variables strategy · CVS · Scientific reasoning · Kindergarten · Dynamic assessment

J. van der Graaf · E. Segers · L. Verhoeven
Behavioural Science Institute, Radboud University Nijmegen, Nijmegen, The Netherlands

J. van der Graaf (✉)
Behavioural Science Institute, Radboud University Nijmegen, Montessorilaan 3, Room A05.16, P.O. Box 9104, 6500 HE Nijmegen, The Netherlands
e-mail: j.vandergraaf@pwo.ru.nl

Scientific reasoning is at the heart of science and technology education, which is becoming more and more important from both international and economic perspectives, because world-wide expenditures on science and science education are increasing (OECD 2013). Young children are curious by nature (Engel 2009) and science and technology are gradually taking their place in kindergarten education where young children are recognized as "natural scientists" (Gopnik 2012). Kindergartners already have some basic understanding of experimentation (Piekny and Maehler 2013). However, when asked how to successively accumulate evidence (Piekny and Maehler 2013) or to predict and explain the outcome of an experiment (Siegler 1976), they do not perform better than predicted by chance. Letting children design experiments themselves also does not improve their understanding of experimentation, which remains low for both 7-year-olds (Chen and Klahr 1999), and 10-year-olds (Schauble 1996). This low performance appears to be due, in part, to a tendency to change multiple variables at once and thereby make it impossible to identify which change has caused an effect (Wilkening and Huber 2004). A more fine-grained analysis of the ability of young children to design unconfounded experiments using multiple variables is thus needed. In the present study, we therefore adapted the so-called "ramp task" of Chen and Klahr (1999) for the dynamic, i.e. interactive, hands-on assessment of young children's ability to design unconfounded experiments. The aim of the present study was to assess the validity of this assessment and to investigate whether kindergartners can use the CVS. Dynamic assessment was chosen, as such assessment provides optimal information on the learning potential of a child (Tzuriel 2000). Tzuriel (2000) refers to dynamic assessment as "an assessment of thinking, perception, learning, and problem solving by an active teaching process aimed at modifying cognitive functioning" (p. 386). This is an instructional method based on the so-called zone of proximal development, which states that under sufficient guidance children can do many things they fail in individually (Vygotsky 1978).

Scientific reasoning

According to Klahr's "Scientific Discovery as Dual Search" model, scientific reasoning can be seen to consist of three cognitive components: hypothesis generation, experimentation, and evidence evaluation (Klahr 2000; Klahr and Dunbar 1988). With respect to the first component, namely hypothesis generation, children appear to develop this capacity around the age of 7 years (Piekny and Maehler 2013). Piekny and Maehler presented cards with depictions of fantasy animals for identification by children. After presenting some examples of a family of animals, the child had to determine whether the last card in the series belonged to that family or not. The examples provided information on what body parts of the animal could be relevant for belonging to a certain family. This appeared to be extremely difficult for 4- to 6-year-olds, and performance increased around the age of 7 years, but significantly more correct hypotheses were only generated around the age of 11 years.

To gain insight into children's capacity for experimentation or the second component of scientific reasoning, they have asked children how to design an experiment. This was done in a task with a single, dichotomous variable by Piekny and Maehler (2013), who asked 4- to 12-year-old children which mouse house (i.e. either one with a small or one with a large opening) they should pick to (a) feed both a small mouse and a large mouse, or (b) find out whether the mouse that went inside to eat was big or small. More than half of the 5-year-

olds correctly chose the house with a large opening for problem (a), while for problem (b) most children incorrectly chose the house with a large opening, because both mice could get in the house with a large opening. This revealed that their investigation in the experimental context, problem (b), was incorrect, but that most kindergartners were nevertheless able to choose the correct setting in a non-experimental context, problem (a).

In other research on children's experimentation, Siegler and Chen (1998) used a balance beam to explore the scientific reasoning of 4- and 5-year-olds. Using actual materials, the children were asked what they thought would happen to the balance beam. Most of the children could predict and explain which side of the balance beam would go down when only the variable weight was investigated. The 4-year-olds performed equally well to the 5-year-olds after 16 trials accompanied by feedback (i.e. showing the child which side of the balance beam goes down when different weights are on the two sides). However when the children had to incorporate a second variable into their predictions and explanations, namely distance, performance dropped sharply. The 4-year-olds generally could not comprehend the combined effects of weight and distance, while a considerable number of the 5-year-olds could not do this even after 16 trials with feedback. These findings show kindergartners to experience difficulties with the identification and encoding of variables in a multivariable experiment. Siegler and Chen (1998) further showed the kindergartners to have problems reproducing the configuration of a balance beam with weights arranged at different distances. When shown a balance beam for 10 s and then asked to reproduce the configuration of the balance beam using an identical balance beam, the children often reproduced the correct numbers of weights per side of the balance beam, but with the incorrect distance from the fulcrum.

For the third and final component of scientific reasoning, namely evidence evaluation, experimental outcomes have to be interpreted. Kindergartners already show a capacity to detect patterns and draw generalizations as revealed by their ability to categorize numerous plants, animals, and artifacts according to their life status (Opfer and Siegler 2004). The kindergartners were asked questions whether the target item was (a) alive, (b) could move towards goals, (c) grew, or (d) needed water. The questions about moving towards a goal, i.e. goal-directed behavior, led the children to infer that both animals and plants are living things, but artifacts are not. Children in this condition, condition b, performed almost equally to those that were explicitly told this fact, condition a, while the children in conditions c and d were less likely to make this inference. These findings suggest that when relevant variables are emphasized, children can detect the effects of the variables and generalize these to other objects. Children can thus evaluate evidence when it forms a pattern that can be detected and used to consider what caused the patterns (i.e. outcome of an experiment). Koerber et al. (2005) have also documented the ability of children as young as 4 years of age to evaluate evidence. The children were asked about a single, dichotomous variable, namely which of the two colors of chewing gum caused bad teeth. No instruction or feedback was provided. Cards depicted children with either bad or good teeth and either color of chewing gum. The number of cards with children with bad teeth varied per color, which led to a conclusive, suggestive, and inconclusive condition. Kindergartners correctly interpreted conclusive and suggestive evidence, but not inconclusive evidence. In other research, Klahr and Chen (2003) showed that kindergartners can even learn to interpret inconclusive evidence correctly when explicit feedback is given on their evaluation of evidence. Explicit feedback, involving explanations of whether the child's response and explanation were correct and why, was compared to implicit feedback, which involved no direct feedback by the experimenters, but—just as in the explicit condition—the full evidence was presented after the response.

Converging evidence thus suggests that young children can learn to generate hypotheses and to evaluate evidence, but that they have difficulties with experimentation. Studies have shown that children can understand some parts of the process of experimentation, but not others. Performance can be low in experimental contexts and when there are multiple variables in the experiment.

Control of variables strategy

In a multivariable experiment one must keep all but the variable of interest constant and thus manipulate a single variable to determine this effect. Inhelder and Piaget (1958) showed "the method of varying a single factor while holding all other things equal" (p. 75) to be present in children's hands-on experimentation with factors possibly affecting the frequency of a pendulum's oscillations. Chen and Klahr (1999) introduced the term "control of variables strategy" (CVS) to refer to this fact that it is not possible to design a multivariable experiment without controlling for the multiplicity of variables. They defined the CVS in procedural and logical terms. Procedurally, it is a method for creating experiments and distinguishing between unconfounded and confounded experiments. In logical terms, CVS includes the ability to make appropriate inferences from outcomes and the understanding of the inherent indeterminacy of confounded experiments. The CVS can, thus, also be used in situations outside the direct experimental context, such as in the process of engineering (Klahr et al. 2007), and problem-solving and decision-making (Mayer et al. 2014). The CVS is one of various learning processes happening within science education, besides other processes such as statistical learning (Saffran 2002) and knowledge acquisition via explicit instruction (Matlen and Klahr 2013).

Chen and Klahr (1999) studied 7- to 10-year-old children's CVS use with hands-on experimentation. Experiments were designed with three types of materials: springs, ramps, and objects that could sink or float (i.e. sinking). Each consisted of four dichotomous variables. Four phases were followed in their intervention study: exploration, training, assessment, and transfer. During the exploration phase, the child was introduced to the task and made two comparisons. Per comparison the child could set all variables, but was asked to investigate the effect of one. Then, children were probed about the comparisons that they made and what they could tell from the outcome. Three training conditions then followed: explicit instruction with probe questions, only probe questions, or no explicit instruction and no probe questions. The explicit instruction involved the explanation of CVS and the rationale underlying it in addition to giving examples of how to make what are commonly called "unconfounded comparisons". The probe questions consisted of asking the child why they made the comparison the way they did and, after the comparison was conducted, if they could tell for sure whether the variable being investigated had made a difference. During the assessment phase, the children designed experiments using the same materials as in the exploration and training phases, with one familiar and one unfamiliar variable. Transfer was assessed 1 week later by having the children perform comparisons with unfamiliar types of materials. The results showed no improvement in CVS use when no instruction was provided. When probe questions were provided, only the CVS use of the 10-year-olds improved slightly between initial exploration and the transfer task. Explicit instruction proved effective. During the assessment phase and thus directly following explicit instruction, all age groups showed improved CVS use compared to initial exploration. For the 9- and 10-year-olds, moreover, this improvement remained during transfer. The 7-year-olds showed no improvement in CVS use at transfer compared to initial

exploration. They performed worse than the older children directly following instruction. Taken together, these results show the youngest group of children, the 7-year-olds, to have difficulties in designing unconfounded experiments with multiple variables. They were capable of learning to use the CVS correctly, but this learning effect was small and short-lived.

In sum, the results of studies of CVS use appear to be inconclusive with regard to young children's ability to use it. These inconclusive findings appear to be related to the design of the studies to date but, nevertheless in light of these studies, four factors can be seen to affect young children's ability to design unconfounded experiments. The first factor is whether the design of the experiment has to be implemented, hands-on, or simply communicated. Although this comparison has not been conducted for one and the same design, it appears that actually having to build the experiment hands-on is more challenging, but also a better approximation of real life situations (e.g., Chen and Klahr 1999) than simply being asked about the design of the experiment (e.g., Piekny and Maehler 2013). A second factor is the identification and encoding of the relevant variables for an experiment, which can be troublesome for young children. When children as old as 7 years are explicitly introduced to the relevant variables, they still show difficulties using CVS (Chen and Klahr 1999). Following a period of feedback, kindergartners still cannot incorporate all variable into their strategy for creating an unconfounded experiment (Siegler and Chen 1998). The third factor affecting their reasoning and design ability is the number of variables that might affect the outcome of the experiment. When kindergartners are asked to set one variable to be able to draw conclusions from the experiment, performance is good (Piekny and Maehler 2013). When they are asked to do this for multiple variables, performance declines (Wilkening and Huber 2004). The fourth factor is instruction. For 10-year-olds, explicit instruction on the CVS in the classroom has been shown to be more effective than having the children design and run their own experiments to study the effect of the variables on the outcome (Lorch et al. 2010). When no explicit instruction is provided, no improvement in the use of CVS is found for 7- to 10-year-olds (Chen and Klahr 1999). Also even explicit instruction has been found to produce only a small but, short-lived improvement in CVS usage among the 7-year-olds (Chen and Klahr 1999). Explicit instruction thus appears to be effective for teaching the CVS to children 10 and older; it does not induce extensive or long-lasting conceptual change in younger children. Younger children might therefore benefit from more dynamic assessment (Tzuriel 2000).

The present study

To help fill some of the gaps in previous research on scientific reasoning in young children, we developed and validated a measure of CVS use by kindergarten children. The aims of the present study were to develop and validate a measure of CVS use and to investigate whether kindergartners can use the CVS. Education tries to incorporate science and technology (OECD 2013), which can be learned via scientific reasoning. Young children are curious (Engel 2009), and they might learn from their exploration when they know the CVS and how to use it. To the best of our knowledge, no studies have been conducted on the hands-on use of the CVS in multivariable experiments with kindergartners. They might nevertheless be able to design multivariable experiments when the four factors outlined above are clearly taken into consideration and an adequate measure of CVS use is employed. To examine CVS use, we therefore adapted one of the experimentation tasks employed by Chen and Klahr (1999) for hand-on use by children aged 4–6 years old.

The four factors that are known to influence the ability to design unconfounded experiments were taken into account in the following manner. First, our version of the ramp task was made hands-on for use with younger children. In such a manner, the design of the task presumably approximates the actual explorations of the natural world by the children. Potential problems with the verbalization of the planned design were also minimized by having the children actually build the experiment. Second, it is known that identification and encoding of relevant variables for purposes of experimentation can be difficult four young children. We therefore explicitly introduced the variables at the start of the dynamic assessment task. This was done by having children interact with them and the experimenter providing the names of the variables. The children were also asked to reproduce the name of the variables. Third, the number of variables might affect performance and we therefore explicitly examined this in the present study. In our version of the ramp task, we started with only one experimental variable to be set. When the children showed an understanding of this in their performance, the number of variables to be set was increased by one. This resulted in a total of four possible levels of experimentation in the end. It also makes the task dynamic. Fourth, explicit instruction of young children has been shown to induce small but short-lived understanding. Feedback was therefore provided following each experiment. When the child's design of an experiment was correct, they were told that it was correct and that they had thus set the relevant variable(s) correctly. When the child's design of an experiment was incorrect, it was explained why the design was incorrect and how it could be done correctly, while the experimenter correctly set the variables.

To validate the dynamic assessment of the children, the content validity was determined in a Mokken scale analysis (MSA) (Mokken 1971). Prior to this validation, the reliability estimates were calculated for the performance measures. The dynamic assessment resulted in four levels, which increased in difficulty, because the number of variables increased. To validate performance on individual levels and over all levels, it had to be indicative of the same underlying trait. MSA validates this by scaling individual levels and forming one total scale of performance across all levels. When the levels are found to be scalable, they can be assumed to measure the same construct, in this case CVS use. The increased difficulty of the levels can also be validated, which the MSA does by ordering the levels along the scale. At the end of the scale levels were more difficult, because more skill was needed to respond correctly. Convergent validity was investigated in terms of nonverbal reasoning, which involves inductive and deductive reasoning. Reasoning has been suggested to be an integral part of scientific reasoning (Zimmerman 2000). According to Dunbar and Klahr (2012) scientific discourse "includes the set of reasoning processes that permeate the field of science: induction, deduction, experimental design, causal reasoning, concept formation, hypothesis testing, and so on". Various types of reasoning seem to be relevant in scientific reasoning. Both nonverbal reasoning and scientific reasoning have been shown to improve with age and relate to each other throughout elementary and middle school (Zimmerman 2007). Following validation, the relations between CVS use, and age and gender were further explored.

Methods

Participants

A total of 46 children from an elementary school in the Netherlands participated. One of them did not perform the ramp task, due to illness. All of the children were in kindergarten,

which is a 2 year program in the Netherlands after which formal education starts. There were 14 girls and 9 boys in the first year (K1), 9 girls and 14 boys in the second year (K2). The average age was 5 years and 3 months, with a range from 4 years and 6 months to 6 years and 3 months. The school was a so-called "talent hotbed school", which means—among other things—extra attention to science and technology during the children's education.

Active consent was given by the parents/caretakers of the children who participated. Upon completion of the study, children received a small reward. Of the parents, 18.5 % had attained an elementary education, 50 % had attained no more than a secondary education, 25 % had attained a tertiary education, and 6.5 % did not report their highest level of education. This indicates that the socioeconomic status of the study participants was slightly above the Dutch averages of 33.8, 39.3 and 17 %, respectively (Eurostat 2013).

Materials

Experimental ramp task

To assess knowledge and use of CVS, we made use of two wooden ramps similar to those used in the study by Chen and Klahr (1999), see Fig. 1. The ramps consisted of a slope and a stepped area. Balls could thus roll down the slope and stop in the stepped area. Four variables could be manipulated to influence how far the ball will roll: ball type (i.e. weight),[1] steepness of the slope (i.e. incline), starting gate (i.e. distance), and surface texture (i.e. friction). Per ramp, thus: a heavy or light ball could be selected; the slope could be made steep or less steep by placing a wooden block under the ramp; the starting gate could be positioned near the top of the slope or further down the slope; and the surface of the slope could be made smooth or rough depending on the choice of plank placed on top of the slope. All variables could affect how far the ball would roll.

Procedure

The ramp task was administered in a single session with an average duration of about 45 min. All children first received instruction and then went on with designing experiments. The children's nonverbal reasoning was assessed in a second session. All of the children were tested individually in a quiet place in the school by the same experimenter, i.e. the first author of this paper.

Introduction of the ramp task

To introduce the task at hand, the children were told that they were going to play a little professor for a while and do some experiments. This was told, because a professor generally is viewed as an authority in science, and thus in experimenting. An example, unrelated to the current experiment, was provided of an experiment. Two drawings of two different pairs of shoes were presented. The children were told that if they wanted to find out which pair of shoes would let them run faster, they could compare the shoes by first running with one pair and then with the other.

The experimenter next introduced the task by setting it up similarly for each child and showing each child how it worked. The child was shown how the ball would roll when the starting gate was opened. The child was then invited to open the gate him/herself to see

[1] The correct term for weight here is mass.

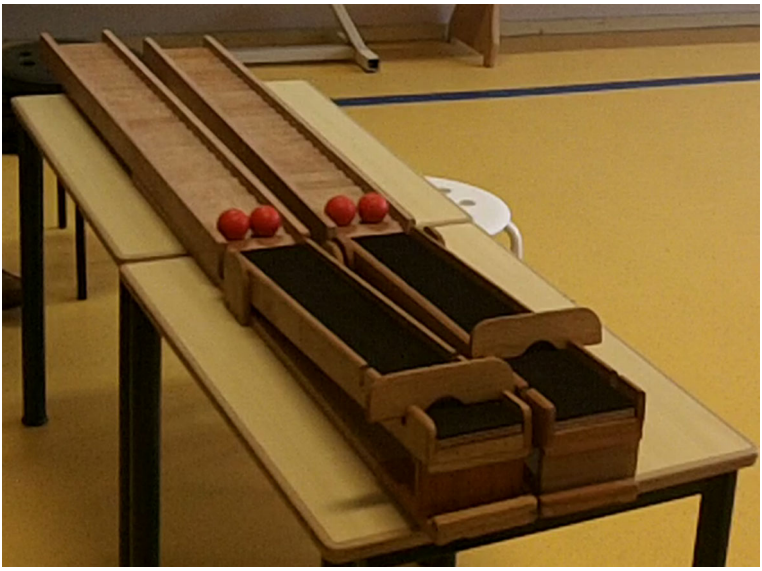

Fig. 1 Photograph of two ramps used in present study side-by-side. The stepped surface to measure how far the ball rolls is depicted on the far side of the table. Both of the ramps depicted here have a rough surface, and steep slopes. One of the starting gates is set near the top of the slope; the other set further down. The balls have yet to be selected. They look the same, but have different weights

how the ball rolled. The experimenter pointed out that it could be determined just how far the ball had rolled by counting the number of steps that the ball had travelled and then did this out loud. The children were further told that the ramps were special, because they could be changed in a number of ways and experiments thus could be done to investigate how far the balls will roll when things have been changed.

Dynamic assessment of CVS

The dynamic assessment of the children's use of CVS consisted of four experiments to be set up by the child, see Fig. 2. To start with the child was asked to experiment with one variable in each of four experiments (Level 1). In each of the Level 1 experiments, a single variable was thus investigated (i.e. one of the four possible variables).

The individual variable was introduced. The children were invited to play around with the variable—for example, by weighing the balls when this was the variable of interest, or touching the different surfaces when this was the variable of interest. The children were asked if they could show which of the balls—the heavier of the lighter—would roll further. For each ramp, a heavy or light ball could be chosen by the child. The effect of the variable was then visually inspected by comparing how far the balls went on the respective ramps. In this example, only the ball was the variable of interest and therefore allowed to be set by the children. The experimenter had already set the other three variables in the same manner for the two ramps.

Once the children had built the ramps the way they wanted to, the experimenter asked them why they had built the ramps the way they did. If they built the ramps correctly, the experimenter told them that they had done this correctly and why. The experimenter might point out, for example, that the variables, that had to be controlled, were correctly set up

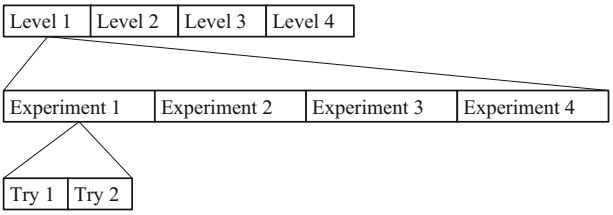

Fig. 2 Schematic overview of the ramp procedure

similarly for the two ramps, while the variable of interest was correctly set up differently. In addition, it might be said that the child's set up was a good set up to study the effects of the variable being investigated. If the child designed the experiment incorrectly, they were told that the design was not yet completely correct and that they should try doing it again, but differently. If, after two tries, the solution was still incorrect, the experimenter explained how the experiment should be designed and meanwhile set up the ramps correctly. The child then let the balls roll and was asked why one ball had rolled further than the other.

Levels 2 through 4 were similar to Level 1, except that the variables were not introduced again. In Level 2, the children were asked to set two variables; in Level 3, three variables; and in Level 4, all four variables. The child could only proceed to the next level when at least one of the four experiments at the current level was designed correctly, either on the first or second try. When all four experiments at one level were designed incorrectly, testing was discontinued. The children could thus be asked to design a total of 16 experiments when at least one experiment was designed correctly at Level 1, 2, and 3. Each child in our study designed a minimum of four experiments (i.e. the four experiments in Level 1).

The administration of the ramp task was arranged to reduce potential confounding and control for possible differences in saliency of the variables. To start with, the order of the variables of investigation was randomized per level. In addition, the first experiment at a level could not investigate the same variable as the last experiment at the previous level.

Scoring of CVS use

Two measures of CVS use were obtained: experiment correct score and variable correct score.

The experiment correct score was defined as the total number of experiments designed correctly, with a maximum of 16 possible points indicating correct design of each experiment on either the first or second try. Each correct design was assigned a score of 1; each incorrect design a score of 0. When testing was discontinued, the unconducted experiments were judged to be incorrect and thus assigned a score of 0. The sum of the correct responses constituted the experiment correct score.

The variable correct score was defined as the total number of variables correctly set for all experiments with a maximum possible score of 40. One point was assigned per correctly set variable (i.e. when the variable of interest was set differently for the two ramps). Additional points were obviously assigned as the number of variables to be set increased at Level 2, 3 and 4. For each variable that was not under investigation, 1 point could be scored when it was kept constant. Only those tries that were correctly set up by the children were used to determine their variable correct scores. When a correct response was not obtained on either of the children's two tries, the setting of the variables on the second try was scored for the children.

Nonverbal reasoning

The children's nonverbal reasoning was measured using an exclusion task (Bleichrodt et al. 1987). This test was included to help us establish the convergent validity of our dynamic assessment of the children's use of the CVS. They were presented four abstract figures and asked to select the one that differed from the other three. Inductive reasoning was needed to determine the underlying category and distinguish category members from non-members. Deductive reasoning was needed to determine the non-member, based on the category. When four consecutive items were responded to incorrectly, testing was discontinued. The child's score was then their correct responses with a total possible score of 30 items (i.e. the total number of items sets presented).

Data analysis

Reliability coefficients were calculated using the Mokken package (Van der Ark 2012) in R (R core team 2014). Reliability was assessed using Cronbach's alpha (Cronbach 1951), Guttman's lambda.2 (Guttman 1945), the Molenaar Sijtsma statistic (Molenaar and Sijtsma 1988), and the latent class reliability statistic (LCRC; Van der Ark et al. 2011). The coefficients were estimates of internal consistency, with values between .60 and .70 judged as acceptable, values between .70 and .80 judged as good, and values larger than .80 judged as excellent in dealing with psychological constructs (Kline 1993).

To assess the content validity of our assessment of CVS use, we employed Mokken scale analysis (MSA) (Mokken 1971). Performance on individual levels and total performance on the ramp task were all scaled. Performance was analyzed in terms of the number of correctly designed experiments per level, because the experiments within a single level could be assumed to be of comparable difficulty. For his analysis, Mokken (1971) proposed the double monotonicity model, which describes relations between and within participants and items. Double refers to the monotony that should manifest itself for both participants and items. Participants were scaled according to their underlying skill (i.e. their CVS use in the present task). Monotony implied that the higher the skillfulness of the CVS use of a participant, the higher the probability of a correct response on an item. In addition to monotony of CVS use, there should also be monotony of item difficulties. The higher the difficulty of an item, the smaller the probability that any given participant will be able to give the correct response. This model dictates four assumptions. The first is unidimensionality, or the assumption that only a single underlying skill is needed to explain the associations between item scores. When this underlying skill is controlled for, the responses to different items should be unrelated, since these relations should be explained by the underlying skill. This is the second assumption of local independence. The third assumption is latent monotonicity, which implies that the item rest functions are globally increasing and not decreasing functions of the underlying skill. Item rest functions reflect the probability of a correct response on an item being a function of the underlying skill. Finally, nonintersection is assumed and thus that the item rest functions should not cross each other. MSA provides four measures, which can be used to determine if the double monotonicity model should be rejected or not. When these assumptions are met, it can be concluded that the model holds and that the present ramp task validly measures CVS use.

MSA was performed in R, version 3.1.0 (R core team 2014), using the Mokken package (Van der Ark 2007; 2012). This package provides four coefficients that relate to aspects of the double monotonicity model and thus indicate whether the model and its underlying assumptions hold. First, the item-pair scalability coefficient, *Hij*, should be positive for

items that belong to the same scale (Mokken 1971). This value was the normed covariance between two item scores when their variances were both positive. It reflects the degree to which the two items vary together and thus can be seen as a measure of the degree to which the items do not overlap in their item step functions. Second, the item scalability coefficient, *Hj*, should be larger than .30 (Mokken 1971). This cut-off is most often used in software (Van der Ark 2012). The item scalability coefficient, *Hj*, measured the association between an individual item and the underlying trait. This coefficient can be interpreted as a discrimination parameter. Third, the test-scalability coefficient, *H*, should be between .30 and .40 for a weak scale, between .40 and .50 for a moderate scale, and larger than .50 for a strong scale (Mokken 1971). This coefficient refers to the degree to which the ordering of participants according to their test scores accurately reflects their ordering according to the underlying trait. Fourth, an item-ordering coefficient, *Ht*, should be between .30 and .40 for weak ordering, between .40 and .50 for moderate ordering, and larger than .50 for strong ordering (Ligtvoet et al. 2010). This value indicated whether the items were ordered on the basis of their difficulty with Level 1 being easiest and Level 4 being most difficult.

Additional analyses were performed in SPSS, Version 19. Children from K1 and K2 were compared using independent samples t-tests. Effect size of potential differences was calculated with Cohen's *d*, which is small when between .20 and .50, medium when between .50 and .80, and large when above .80 (Cohen 1988). Item ordering by MSA was confirmed with a repeated-measures ANOVA with level as a within-subjects factor and gender as a between-subjects factor. Correlations were calculated between nonverbal reasoning and CVS use, with a Pearson's *r* between .10 and .30 being small, between .30 and .50 being medium, and above .50 being large (Cohen 1992). Finally, linear regressions were performed with age and gender entered simultaneously to determine their relation to CVS use.

Results

Reliability

The reliability of the experiment and variable correct scores across 16 experiments and four levels was found to be acceptable to excellent, see Table 1. The Molenaar Sijtsma statistic and the LCRC provided suitable estimates given current data, lambda.2 is a good alternative, while Cronbach's alpha can be less accurate and more biased (Van der Ark et al. 2011). The Molenaar Sijtsma statistic and the LCRC showed reliability to be good to

Table 1 Reliability Coefficients

Variable	Scale	Molenaar Sijtsma statistic	LCRC	Cronbach's alpha	Lambda.2
Experiment correct score	Levels	.79	.75	.68	.74
	Experiments	.87	.93	.70	.76
Variable correct score	Levels	.83	[a]	.74	.79
	Experiments	.75	[a]	.73	.78

[a] Latent class reliability statistic (LCRC) was not calculated for variable correct score. The Mokken model did not hold for the variable correct score, as the maximum score increased with level

excellent, Lambda.2 showed it to be good and Cronbach's alpha showed it to be acceptable (Kline 1993).

Content validity

For MSA levels were taken as items. The total number of correctly designed experiments per Level was averaged across participants, see Table 2. This score per level was then plotted against the number of correctly designed experiments for the other levels considered together and referred to as the experiment correct rest score; see Fig. 3. As can be seen, the slope of Level 1 increased first, followed by the slope of Level 2. The slopes of Levels 3 and 4 were found to be comparable. These findings show Level 1 to be easiest, followed by Level 2, and then by the more difficult Level 3 and 4. The item rest functions appeared to be generally increasing, which means that when someone showed more skilled CVS use, they also correctly designed more experiments at each level.

The general increase in the item-rest functions was further investigated by checking if the double monotonicity model held with its four assumptions. This was done with the four coefficients provided by the MSA. The first coefficient was the item-pair scalability coefficient, *Hij*. Confidence intervals were calculated, see Table 2. The confidence interval should not include zero for the coefficient to be significantly larger than zero. Given that this was a one-tailed test, as recommended by Mokken (1971), a confidence interval of 90 % was used. Three out of six item-pair coefficients showed a lower bound around zero, which suggests nonsignificance. There was one violation between the pair of Levels 1 and 4. The item-pair coefficient was larger than zero, $Hij = .34$, but the standard error was also large, $SE = .32$, which can be due to two factors. The standard errors of *Hij* coefficients are generally quite large (Van der Ark 2012). The relatively small sample size of 45 participants and/or the relatively changeable covariance between participants, due to the large difference in difficulty between the levels, could have contributed to the large standard error. For these reasons, all of the item-pair coefficients were assumed to be larger than zero.

Second, the item scalability coefficient, *Hj*, was investigated per Level and should be larger than .30. The value of .30 has been proposed as the cut-off point for the absolute value of the item scalability coefficient for the item to provide information on the underlying skill by discriminating between participants (Mokken 1971). In addition, 90 % confidence intervals were calculated. The item scalability coefficients per Level showed the levels to discriminate between participants according to their CVS use, see Table 3. Level 2 showed the lowest discrimination parameter and Levels 3 and 4 the highest. Although Levels 1 and 2 included .30 in their confidence intervals, the item scalability coefficients were larger than .30 and significantly larger than zero. All of the levels therefore provided satisfactory item scalability coefficients.

Table 2 Item-pair coefficient, Hij, between the levels

Levels	1		2		3	
	Hij	90 % CI	*Hij*	90 % CI	*Hij*	90 % CI
1	–					
2	.48	[.24, .71]	–			
3	.30	[.00, .61]	.26	[-.02, .54]	–	
4	.34	[-.19, .87]	.29	[-.01, .57]	.91	[.84, .97]

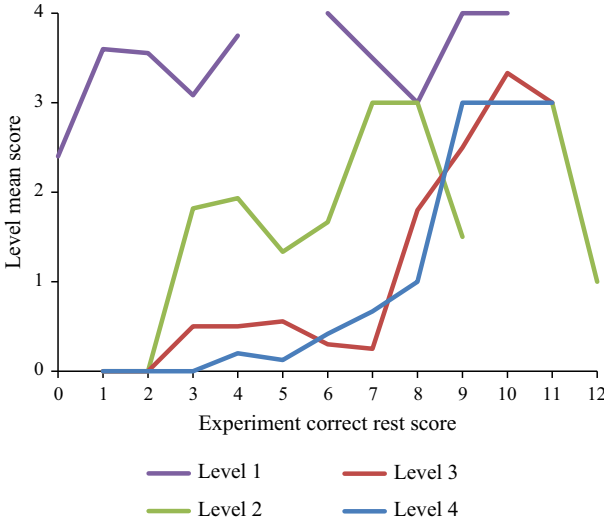

Fig. 3 The item rest functions per level (i.e. average experiment correct score per level as a function of experiment correct rest score)

The third coefficient, namely the test-scalability coefficient, *H*, showed the ordering of the participants according to their test score (i.e. experiment correct score), to accurately reflect the ordering of participants on the underlying skill (i.e. CVS use). It was assumed that CVS use was reflected by the experiment correct score, which was the sum of all correctly designed experiments, and that all of the levels measured the same skill and thus form a single scale. Levels were partitioned into a single Mokken scale with an automated item selection procedure used to do this. The genetic algorithm was used because it has been claimed to be a better algorithm than the hierarchical clustering algorithm (for details, see Straat et al. 2013). The test-scalability coefficient, $H = .48$, was in the moderate range, which shows the ordering of the participants by test score to accurately reflect their ordering according to the underlying trait. The lower bound of the 90 % confidence interval was .32, which shows the scale to be weak at minimum.

Finally, the item-ordering coefficient, *Ht*, showed the levels to be ordered accurately and difficulty thus to increase with Level. The *Ht* indicates weak ordering between .30 and .40, moderate ordering when between .40 and .50, and strong ordering when larger than .50 (Mokken 1971). When the order of the levels was investigated using the manifest invariant item ordering (MIIO) method, which is a method for analyzing polytomous items, the levels were shown to be ordered accurately and subsequent levels to be more difficult, $Ht = .84$. The item-ordering coefficient indicated strong ordering. Whether performance dropped as difficulty, expressed in Level, increased was further investigated. A repeated-

Table 3 Mean experiment correct score, item scalability coefficient, Hj, as a function of level

Level	*M (SD)*	*Hj*	90 % CI
1	3.36 (.83)	.38	[.11, .65]
2	1.84 (1.07)	.33	[.10, .56]
3	.89 (1.21)	.57	[.43, .71]
4	.73 (1.26)	.60	[.44, .76]
Total	6.82 (3.17)	.48	[.32, .65]

measures ANOVA with Level as within-subjects factor and gender as between-subjects factor was conducted. Gender was added to detect potential interactions with Level, but it had no effect. There was a significant main effect of Level, $F(2.07, 89.05) = 80.46$, $p < .001$. Note that sphericity could not be assumed, Mauchly's $W(5) = .40$, $p < .001$. Therefore, the Huyhn-Feldt correction was applied (Huynh and Feldt 1976). The number of correctly designed experiments decreased linearly with level, $F(1,43) = 127.27$, $p < .001$, see Table 3. Post-hoc analyses revealed that performance was higher on Level 1 than on any other level; and performance was higher on Level 2 than on Levels 3 and 4, $p < .001$, while performance on Levels 3 and 4 did not differ significantly. This latter finding is in keeping with the comparable slopes for the item rest functions of Levels 3 and 4, see Fig. 3.

The values of the four MSA coefficients were sufficient to justify the conclusion that the model's assumptions were met. This means that the double monotonicity model held for the data collected and that content validity has been established for our version of the ramp task.

CVS use as a function of nonverbal reasoning, age, and gender

Having shown the children's CVS use to be measured reliably and validly, their performance on the ramp task was further analyzed. All of the kindergartners designed at least one experiment correctly at Level 1 and were thus able to proceed to Level 2. Out of the total of 45 children, 40 were able to correctly design at least one experiment with two variables (i.e. at Level 2) and thus proceeded to Level 3. All 23 children from K2 (i.e. the second year of kindergarten) were able to do this, while 18 out of 23 children from K1 (i.e. the first year of kindergarten) were able to do this. At Level 3, 21 of the children were able to correctly design an experiment with three variables and thus proceeded to Level 4, where 14 out of 21 children correctly designed at least one experiment with four variables and 7 designed all of the experiments with four variables incorrectly.

We next compared the different years of kindergarten (i.e. K1 vs. K2). The children in K2 showed better CVS scores (i.e. experiment and variable correct score) and nonverbal reasoning scores than the children in K1, see Table 4. The differences in the K1 versus K2 children's CVS scores were large, as indicated by a Cohen's d of 1.21 for experiment correct scores and a Cohen's d of 1.22 for variable correct scores, while nonverbal reasoning scores showed a medium effect size (Cohen, 1988).

Convergent validity of the children's CVS scores was assessed by relating these to their nonverbal reasoning scores. Positive, medium correlations (Cohen 1992) were found for

Table 4 Descriptives and contrasts between K1 and K2

	K1		K2		p	Cohen's d
	M	*SD*	*M*	*SD*		
Nonverbal reasoning	17.43	4.93	20.35	4.81	.049	.60
CVS scores						
Experiment correct	5.14	2.51	8.44	2.92	<.001	1.21
Variable correct	14.68	8.12	24.65	8.18	<.001	1.22

the children's nonverbal reasoning scores with their experiment correct scores, $r(43) = .47$, $p = .001$, and with their variable correct scores, $r(43) = .42$, $p = .004$. The extent to which children's performance related to their age and/or gender was also explored. Linear regression models were built with gender and age in months simultaneously entered as independent variables and the experiment or variable correct scores used as dependent variable. Age related positively to both the experiment correct scores, $\beta = .51$, $p < .001$, and the variable correct scores, $\beta = .47$, $p = .002$. Gender did not relate to either the experiment correct scores, $p = .469$, or variable correct scores, $p = .314$.

Discussion

The aim of the present study was to validate a newly developed, dynamic assessment approach to the measurement of CVS use by kindergartners' ages 4–6 years, and to investigate whether kindergartners can use the CVS. According to different procedural measures, the assessment was found to be reliable. To address its content validity, we examined the scalability of the participants and items in a Mokken scale analysis (MSA; Mokken 1971). The MSA coefficients were sufficient to conclude that the double monotonicity model held, which is fully commensurate with a unidimensional interpretation of the dynamic assessment of CVS.

The four levels indeed formed a single scale with the levels accurately ordered according to difficulty (i.e. Level 1 easiest, Level 4 most difficult). The order confirmed the aims and design of the task as the number of variables that the children had to set increased per level. Using CVS thus became more difficult as the number of variables increased. This is in line with the results of previous research (e.g., Siegler and Chen 1998) and shows that the number of variables that the children are required to set can be used to scale them according to their CVS use. The most skilled children should respond correctly on all levels, while the least skilled children should respond correctly on only Levels 1 and 2, but not on Levels 3 and 4.

The individual levels also had satisfactory discrimination parameters, which means that the children could be differentiated according to their CVS use on a single level. While the discriminatory parameters could not be compared statistically, visual inspection of them showed Levels 3 and 4 to have larger discrimination parameters than Levels 1 and 2. Experiments with three or four variables to be manipulated are thus good indicators of CVS use by kindergartners (i.e. 4–6 years of age).

To demonstrate the validity of our dynamic assessment of CVS use, we examined its association with the nonverbal reasoning of the children. We found nonverbal reasoning to be positively related to CVS use, which indicates that nonverbal reasoning is involved in learning and applying CVS at the kindergarten level. Given that nonverbal reasoning can be considered part of scientific reasoning (Klahr 2000; Zimmerman 2000), the present positive correlation is taken to be an indicator of the convergent validity of our dynamic assessment method.

CVS use was further found to be related to age but not gender. The positive relation to age can be interpreted in terms of development. Usually children's performance increases with age, such as on cognitive tasks (Steinberg 2005) and scientific reasoning tasks (Zimmerman 2007). Therefore this relation was assumed to support the validity of present ramp task. In line with this result, children in K2 scored higher than children in K1, which was a large difference. It is therefore possible that age and/or year in school underlies the

development in scientific reasoning. In other words, improvements in scientific reasoning might be due to biological development and/or due to the amount of education given at school. The present study showed that both affect the use of CVS, but to disentangle the underlying factor of the development, future research can investigate possible moderation of age effects by grade. It might be possible that age effects are larger in K1 than in K2, because K1 is the first year of education.

The finding of no relation of gender with CVS use may be due to young age of the children in the present study. In older children, gender has been shown to relate to aspects of scientific reasoning, including the following: nonverbal reasoning and science school grades (Kuhn and Holling 2009), experience with science, interest in science, and attitudes towards science (Jones et al. 2000). Boys have been found to score generally higher on these aspects than girls. A different explanation might be that possible gender differences were attenuated by the dynamic assessment. Dynamic assessment has been shown to eliminate differences on test performance, for example between children with low and high SES on a complex problem-solving task with abstract problems (Tzuriel 2000).

The present results show that CVS use can be measured in kindergarten. As the number of variables to be manipulated increased, the difficulty of the experiments increased. The number of variables to be manipulated can thus be used to scale children's CVS use with different items. The total number of experiments designed correctly demonstrated an early capacity for CVS use and thus show scientific reasoning to already be present in kindergarten. Its development, moreover, can be reliably measured using the dynamic assessment method.

Of course, several limitations apply to the present study. To begin with, the sample size was relatively small, which can be troublesome for MSA. This can be revealed by the calculation of standard errors for the coefficients, which is part of the Mokken package (Van der Ark 2012). The present results showed all of the coefficients to be larger than the cut-off points proposed by Mokken (1971) but, in a few cases, not significantly larger than the cut-off. The results of the MSA should therefore be interpreted with caution and replication should be sought in the future. Another limitation is the ecological validity, because the experimental conditions differ from school settings. While the present study assessed children in an individual setting, it would be valuable to observe children's scientific discourse in an open scientific learning environment. It would furthermore be interesting to investigate the saliency of experimental variables, by allowing children to choose freely between variables, and how that might affect the design of the experiment and/or the conclusions drawn from the evidence.

It can be suggested that the dynamic assessment of the CVS can be expanded to younger and older samples, because all children were successful at Level 1 and some failed at Level 4. It also remains to be investigated whether age and/or grade underlies the development in CVS use. The effects of age were investigated cross-sectionally in this study. Additional longitudinal study can shed light on the course of development for CVS use. It would be interesting to study individual differences in the course of development in relation to other skills kindergartners already possess. Nonverbal reasoning, as a critical part of scientific reasoning, should certainly be assessed in any longitudinal study (Zimmerman 2007). Other skills might be included to gain insight into their role in CVS use and the development of scientific reasoning. Working memory is a good candidate due to its role in learning potential (Paas et al. 2003) and the limitations that it is known to impose on young children's learning (Gathercole et al. 2004). Other candidates for inclusion as variables in relation to CVS use in future studies are other aspects of scientific reasoning such as hypothesis generation (Van Joolingen and De Jong 1991) and evidence evaluation (Metz

2011). And other topics for consideration are the role of science education and the development of scientific knowledge (Driver et al. 1994), a science vocabulary (Leung 2008), and attitudes towards science (Kuhn and Holling 2009).

With respect to actual educational practice, note should be taken of the feasibility demonstrated in present study of dynamically assessing kindergartners' CVS use. This finding is in line with the work of Vygotsky (1978) who has shown that children can do many things that fall within their so-called zone of proximal development when given sufficient guidance—things that they will be able to do on their own a few years later. Kindergartners can thus be exposed to hands-on, multivariable experiments and helped to explore them with the guidance of a teacher or other students. When they are older, they can presumably do this on their own. Given that the dynamic assessment task used in this study showed the children to be capable of correctly designing multivariable experiments with up to four variables manipulated at times, it can be recommended that kindergartners be exposed to such multivariable tasks and experiments. This can nicely prepare them for their further science education in which knowledge often is based on experiments with multiple variables.

The CVS can presumably be taught, and we have shown that dynamic assessment can be used, as an instructional method, to teach it in a single session. Research on CVS use has further shown that direct, explicit instruction results in better knowledge and use of the strategy than unstructured exploration and learning (Klahr et al. 2011; Lorch et al. 2010). Simply structuring a task has been shown to help children's experimentation and inferencing, but have little effect on their mastery of the CVS (Lazonder and Egberink 2014). In the present study we therefore introduced a new procedure for structuring a task that measures CVS use, and this was found to be effective for teaching the CVS. For the teaching of the CVS and perhaps scientific reasoning in general, the structure of a task, thus, appears to be critical. Teaching with only invalid CVS experiments has recently been shown to be more effective than with only valid experiments (Lorch et al. 2014). While these findings still need to be confirmed, previous research has shown preschoolers to engage in more exploratory play when exposed to confounded as opposed to unconfounded evidence (Schulz and Bonawitz 2007). Or, stated differently, what Schulz and Bonawitz refer to as "serious fun" often entails a search for the causal structure underlying observed evidence (i.e. scientific discovery). The CVS can help in this scientific discovery and it appears to be involved in scientific topics in education, such as biology and chemistry. The CVS can also be applied in many more life situations that involve changes in multiple variables, such as understanding social phenomena and in the process of decision-making.

Taken together, these results suggest that science education programs can now be evaluated with regard to their capacity to produce an understanding of the CVS and its correct usage. This aspect of scientific reasoning is particularly relevant when scientific reasoning is conceptualized as involving the intentional seeking of knowledge via application of the methods of scientific inquiry (Kuhn 2004). Children can gain knowledge from their own exploration and experiments with the application of the CVS as part of these. Together with the knowledge that young children are curious by nature (Engel 2009) and the knowledge that hands-on experimentation in the classroom can foster interest in science both inside and outside the school (Ornstein 2005), we can conclude that even young children can and should, thus, be encouraged by schools to explore how the physical world works.

To conclude, the present dynamic assessment of CVS use by kindergartners was valid and reliable. In addition, this method proved to be effective in teaching young children to

use CVS. In this regard, the present study is one of the first to show that children as young as kindergartners can use the CVS and design unconfounded experiments correctly.

Acknowledgments This research was funded by a grant from the Dutch Curious Minds research program. The authors would like to thank David Klahr and his colleagues for kindly providing us with the ramp task and sketches of the materials, and Huub Hennissen for building the ramps. We thank Eddie Dennessen and Juul Ellis for their statistical advice.

References

Bleichrodt, N., Drenth, P. J. D., Zaal, J. N., & Resing, W. C. M. (1987). *Revisie Amsterdamse kinder intelligentie test. Handleiding*. [Revision Amsterdam Child Intelligence Test. Manual]. Lisse: Swets & Zeitlinger.

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155–159. doi:10.1037/0033-2909.112.1.155.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297–334. doi:10.1007/BF02310555.

Driver, R., Asoko, H., Leach, J., Scott, P., & Mortimer, E. (1994). Constructing scientific knowledge in the classroom. *Educational Researcher, 23*, 5–12. doi:10.3102/0013189X023007005.

Dunbar, K. N., & Klahr, D. (2012). Scientific thinking and reasoning. In K. J. Holyoak & R. G. Morrison (Eds.), *The Oxford handbook of thinking and reasoning*. Oxford Handbooks Online. doi:10.1093/oxfordhb/9780199734689.013.0035.

Engel, S. (2009). Is curiosity vanishing? *Journal of the American Academy of Child and Adolescent Psychiatry, 48*, 777–779. doi:10.1097/CHI.0b013e3181aa03b0.

Eurostat, (2013). *European social statistics*. Luxembourg: Publications Office of the European Union. doi:10.2785/36105.

Gathercole, S. E., Pickering, S. J., Ambridge, B., & Wearing, H. (2004). The structure of working memory from 4 to 15 years of age. *Developmental Psychology, 40*, 177–190. doi:10.1037/0012-1649.40.2.177.

Gopnik, A. (2012). Scientific thinking in young children: Theoretical advances, empirical research and policy implications. *Science, 337*, 1623–1627. doi:10.1126/science.1223416.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika, 10*, 255–282. doi:10.1007/BF02288892.

Huynh, H., & Feldt, L. S. (1976). Estimation of the Box correction for degrees of freedom from sample data in randomized block and split-plot designs. *Journal of Educational and Behavioral Statistics, 1*, 69–82. doi:10.3102/10769986001001069.

Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence: An essay on the construction of formal operational structures* (A. Parsons & S. Milgram, Trans.). New York, NY: Basic Books.

Jones, M. G., Howe, A., & Rua, M. J. (2000). Gender differences in students' experiences, interests, and attitudes towards science and scientists. *Science Education, 84*, 180–192.

Klahr, D. (2000). *Exploring science: The cognition and development of discovery processes*. Cambridge, MA: MIT Press.

Klahr, D., & Chen, Z. (2003). Overcoming the positive-capture strategy in young children: Learning about indeterminacy. *Child Development, 74*, 1275–1296.

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

Klahr, D., Triona, L. M., & Williams, C. (2007). Hands on what? The relative effectiveness of physical versus virtual materials in engineering design project by middle school children. *Journal of Research in Science Teaching, 44*, 183–203. doi:10.1002/tea.20152.

Klahr, D., Zimmerman, C., & Jirout, J. (2011). Educational interventions to advance children's scientific thinking. *Science, 333*, 971–975. doi:10.1126/science.1204528.

Kline, P. (1993). *The handbook of psychological testing*. London: Routledge.

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*, 141–152. doi:10.1024/1421-0185.64.3.141.

Kuhn, D. (2004). What is scientific thinking and how does it develop?. In U. Goswami (Ed.), *The blackwell handbook of childhood cognitive development*. Blackwell Reference Online. doi:10.1111/b.9780631218418.2004.00020.x.

Kuhn, J., & Holling, H. (2009). Gender, reasoning ability, and scholastic achievement: A multilevel mediation analysis. *Learning and Individual Differences, 19*, 229–233. doi:10.1016/j.lindif.2008.11.007.

Lazonder, A. W., & Egberink, A. (2014). Children's acquisition and use of the control-of-variables strategy: Effects of explicit and implicit instructional guidance. *Instructional Science, 42*, 291–304. doi:10.1007/s11251-013-9284-3.

Leung, C. B. (2008). Preschoolers' acquisition of scientific vocabulary through repeated read-aloud events, retellings, and hands-on science activities. *Reading Psychology, 29*, 165–193. doi:10.1080/02702710801964090.

Ligtvoet, R., Van der Ark, L. A., Te Marvelde, J. M., & Sijtsma, K. (2010). Investigating an invariant item ordering for polytomously scored items. *Educational and Psychological Measurement, 70*, 578–595. doi:10.1177/0013164409355697.

Lorch, R. F, Jr, Lorch, E. P., Calderhead, W. J., Dunlap, E. E., Hodell, E. C., & Freer, B. D. (2010). Learning the control of variables strategy in higher and lower achieving classrooms: Contributions of explicit instruction and experimentation. *Journal of Educational Psychology, 102*, 90–101. doi:10.1037/a0017972.

Lorch, R. F, Jr, Lorch, E. P., Freer, B. D., Dunlap, E. E., & Hodell, E. C. (2014). Using valid and invalid experimental designs to teach the control of variables strategy in higher and lower achieving classrooms. *Journal of Educational Psychology, 106*, 18–35.

Matlen, B. J., & Klahr, D. (2013). Sequential effects of high and low instructional guidance on children's acquisition of experimental skills: Is it all in the timing? *Instructional Science, 41*, 621–634. doi:10.1007/s11251-012-9248-z.

Mayer, D., Sodian, B., Koerber, S., & Schwippert, K. (2014). Scientific reasoning in elementary school children: Assessment and relations with cognitive abilities. *Learning and Instruction, 29*, 43–55. doi:10.1016/j.learninstruc.2013.07.005.

Metz, K. E. (2011). Disentangling robust developmental constraints from the instructionally mutable: Young children's epistemic reasoning about a study of their own design. *The Journal of the Learning Sciences, 20*, 50–110. doi:10.1080/10508406.2011.529325.

Mokken, R. J. (1971). *A theory and procedure of scale analysis*. The Hague: Mouton & Co.

Molenaar, I.W., & Sijtsma, K. (1988). Mokken's approach to reliability estimation extended to multicategory items. *Kwantitatieve Methoden, 9*, 115–126. Retrieved from http://arno.uvt.nl/show.cgi?fid=81058.

OECD (2013*). OECD science, technology and industry scoreboard 2013: Innovation for growth*. OECD Publishing. doi:10.1787/sti_scoreboard-2013-en.

Opfer, J. E., & Siegler, R. S. (2004). Revisiting preschoolers' living things concept: A microgenetic analysis of conceptual change in basic biology. *Cognitive Psychology, 49*, 301–332. doi:10.1016/j.cogpsych.2004.01.002.

Ornstein, A. (2005). The frequency of hands-on experimentation and student attitudes towards science: A statistically significant relation. *Journal of Science Education and Technology, 15*, 285–297. doi:10.1007/s10956-006-9015-5.

Paas, F., Renkl, A., & Sweller, J. (2003). Cognitive load theory and instructional design: Recent developments. *Educational Psychologist, 38*, 1–4. doi:10.1207/S15326985EP3801_1.

Piekny, J., & Maehler, C. (2013). Scientific reasoning in early and middle childhood: The development of domain-general evidence evaluation, experimentation, and hypothesis generation skills. *British Journal of Developmental Psychology, 31*, 153–179. doi:10.1111/j.2044-835X.2012.02082.x.

Saffran, J. R. (2002). Constraints on statistical learning. *Journal of Memory and Language, 47*, 172–196. doi:10.1006/jmla.2001.2839.

Schauble, L. (1996). The development of scientific reasoning in knowledge-rich contexts. *Developmental Psychology, 32*, 102–119. doi:10.1037/0012-1649.32.1.102.

Schulz, L. E., & Bonawitz, E. B. (2007). Serious fun: preschoolers engage in more exploratory play when evidence is confounded. *Developmental Psychology, 43*, 1045–1050. doi:10.1037/0012-1649.43.4.1045.

Siegler, R. S. (1976). Three aspects of cognitive development. *Cognitive Psychology, 8*, 481–520. doi:10.1016/0010-0285(76)90016-5.

Siegler, R. S., & Chen, Z. (1998). Developmental differences in rule learning: A microgenetic analysis. *Cognitive Psychology, 36*, 273–310. doi:10.1006/cogp.1998.0686.

Steinberg, L. (2005). Cognitive and affective development in adolescence. *TRENDS in Cognitive Sciences, 3*, 69–74. doi:10.1016/j.tics.2004.12.005.

Straat, J. H., Van der Ark, L. A., & Sijtsma, K. (2013). Comparing optimization algorithms for item selection in Mokken scale analysis. *Journal of Classification, 30*, 75–99. doi:10.1007/s00357-013-9122-y.

R Core Team (2014). *R: A language and environment for statistical computing*. Vienna: R foundation for statistical computing. Retrieved from http://www.R-project.org.

Tzuriel, D. (2000). Dynamic assessment of young children: Education and intervention perspectives. *Educational Psychology Review, 12*, 385–435. doi:10.1023/A:1009032414088.

Van der Ark, L.A. (2007). Mokken scale analysis in R. *Journal of Statistical Software, 20*, 1–19. Retrieved from http://www.jstatsoft.org/v20/a11/paper.

Van der Ark, L.A. (2012). New developments in Mokken scale analysis in R. *Journal of Statistical Software, 48*, 1–27. Retrieved from http://www.jstatsoft.org/v48/i05/paper.

Van der Ark, L. A., Van der Palm, D. W., & Sijtsma, K. (2011). A latent class approach to estimating test-score reliability. *Applied Psychological Measurement, 35*, 380–392. doi:10.1177/0146621610392911.

Van Joolingen, W. R., & De Jong, T. (1991). Supporting hypothesis generation by learners exploring an interactive computer simulation. *Instructional Science, 20*, 389–404.

Vygotsky, L.S. (1978). *Mind in society: The development of higher psychological processes* (M. Cole, V. John-Steiner, S. Scribner, & E. Souberman, Trans. and Eds.). Cambridge, MA: Harvard University Press.

Wilkening, F., & Huber, S. (2004). Children's intuitive physics. In U. Goswami (Ed.), *The blackwell handbook of childhood cognitive development*. Blackwell Reference Online. doi:10.1111/b.9780631218418.2004.00019.x.

Zimmerman, C. (2000). The development of scientific reasoning skills. *Developmental Review, 20*, 99–149. doi:10.1006/drev.1999.0497.

Zimmerman, C. (2007). The development of scientific thinking skills in elementary and middle school. *Developmental Review, 27*, 172–223. doi:10.1016/j.dr.2006.12.001.