International Journal of Science Education

ISSN: 0950-0693 (Print) 1464-5289 (Online) Journal homepage: https://www.tandfonline.com/loi/tsed20

A combined approach to strengthen children's scientific thinking: direct instruction on scientific reasoning and training of teacher's verbal support

Joep van der Graaf, Eva van de Sande, Martine Gijsel & Eliane Segers

To cite this article: Joep van der Graaf, Eva van de Sande, Martine Gijsel & Eliane Segers (2019): A combined approach to strengthen children's scientific thinking: direct instruction on scientific reasoning and training of teacher's verbal support, International Journal of Science Education, DOI: 10.1080/09500693.2019.1594442

To link to this article: https://doi.org/10.1080/09500693.2019.1594442

© 2019 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group

Published online: 28 Mar 2019.

Submit your article to this journal

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=tsed20

INTERNATIONAL JOURNAL OF SCIENCE EDUCATION
https://doi.org/10.1080/09500693.2019.1594442

Routledge
Taylor & Francis Group

OPEN ACCESS

A combined approach to strengthen children's scientific thinking: direct instruction on scientific reasoning and training of teacher's verbal support

Joep van der Graaf[a,b], Eva van de Sande[a], Martine Gijsel[c] and Eliane Segers[a,b]

[a]Behavioural Science Institute, Radboud University, Nijmegen, Netherlands; [b]Department of Instructional Technology, University of Twente, Enschede, Netherlands; [c]Academy of Pedagogy and Education, Saxion University of Applied Sciences, Enschede, Netherlands

ABSTRACT

Inquiry-based lessons have been demonstrated to improve children's scientific thinking (i.e. reasoning abilities and domain-specific knowledge). Although empirical evidence shows that inquiry-based learning requires instruction, research comes from two approaches that have not been bridged yet: direct instruction of scientific reasoning and teacher training of verbal support. We investigated how these two types of instruction separately or combined strengthened children's scientific thinking by comparing four conditions: baseline, direct instruction, verbal support, and a combined approach. Effectiveness of an inquiry-based lesson series on scientific reasoning abilities, vocabulary, and domain-specific knowledge (near and far transfer) were studied among 301 fourth graders. Results showed that both approaches strengthened different components of scientific reasoning abilities, and that a combination of instructions was most effective for scientific reasoning abilities, vocabulary, and domain-specific knowledge. Domain-specific knowledge acquisition was strengthened only when both instructions were provided. It can thus be concluded that each type of instruction has unique contributions to children's science learning and that these instructions complement each other. Our study thus showed that inquiry-based lesson series when preceded by direct instruction of scientific reasoning and scaffolded with verbal support are most effective.

ARTICLE HISTORY
Received 10 September 2018
Accepted 9 March 2019

KEYWORDS
Scientific thinking; inquiry learning; direct instruction; verbal support; primary school; intervention

Under the umbrella term of twenty-first-century skills, the development *scientific thinking* is considered a major goal in STEM (science, technology, engineering, and mathematics) education (Fischer et al., 2014). Scientific thinking combines the generic ability of *scientific reasoning* (i.e. the stages in scientific inquiry: formulate questions, design experiments, draw inferences and conclusions) with having *domain-specific knowledge* about science (Klahr, Zimmerman, & Jirout, 2011). In particular, scientific reasoning is necessary for

CONTACT Joep van der Graaf ✉ j.vandergraaf@utwente.nl Department of Instructional Technology, University of Twente, De Zul 10, Room B225, P.O. Box 217, 7500 AE Enschede, Netherlands

© 2019 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (http://creativecommons.org/licenses/by-nc-nd/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.

children to acquire domain-specific knowledge about STEM through inquiry (Andersen & Garcia-Mila, 2017). Inquiry learning is a common way of teaching STEM in the classrooms (Organisation for Economic Co-operation and Development [OECD], 2014), and research has consistently shown that teacher guidance in inquiry-based learning is essential (Alfieri, Brooks, Aldrich, & Tenenbaum, 2011; Kirschner, Sweller, & Clark, 2006; Lazonder & Harmsen, 2016). However, there is no real consensus on how to best implement teacher guidance in strengthening children's scientific thinking. Direct instruction has been advocated, so that children are explicitly thought how to design and conduct experiments (Dunbar & Klahr, 2012). Another approach is that teachers support the academic language that fosters the formulation of questions and drawing inferences and conclusions (e.g. Barber, Catz, & Arya, 2006). Helping children to optimally gain from inquiry learning may best be done by combining the best of both worlds, which up till now have been studied separately. In the present study, we investigated how an inquiry-based lesson series' effectiveness on children's scientific thinking could best be improved; by implementing a direct instruction, verbal support, or a combined approach in comparison to baseline.

1. Direct instruction of scientific reasoning abilities

The core abilities that underlie scientific reasoning are in line with the overall stages of the inquiry cycle: Hypothesis construction, experimentation, and generating conclusions based on evidence (Klahr, 2000). Hypothesis construction is the process of creating a hypothesis from a research question, which can be created through various reasoning processes, such as analogical and deductive reasoning to move from theory to testable hypotheses. Experimentation consists of designing and conducting experiments in order to obtain evidence to answer the research question. Conclusions can be generated from the evidence through logical and inductive reasoning.

Experimentation is the component of scientific reasoning that is most likely to provide children with the building blocks to apply effective inquiry-based learning (Zimmerman, 2007). An example is that children with better experimentation abilities are better in acquiring domain-specific knowledge through inquiry (Edelsbrunner, Schalk, Schumacher, & Stern, 2015). Experimentation is central to inquiry learning, because with setting up and conducting experiments, children also learn how to formulate questions that are investigable, and how to draw conclusions from the evidence of the experiments. The strategy to design unconfounded experiment with multiple variables is called the *Control of Variables Strategy (CVS)* (Chen & Klahr, 1999). This strategy states that in order to investigate a single variable (the focal variable), one has to manipulate that variable while leaving the other variables constant (i.e. the control variables). Without instruction, children tend to manipulate multiple variables at once (Wilkening & Huber, 2004). This is likely due to their tendency to generate effects, which are often in line with their own expectations (Dunbar & Klahr, 2012).

It has been shown that CVS skills can improve in children of the upper primary grades after a single short training. Children that received this instruction showed stronger improvements in inquiry abilities compared to a group of children in a free exploration condition, where no instruction on experimentation skills were given (Wagensveld, Segers, Kleemans, & Verhoeven, 2015). This result has been established consistently, as shown by a review of 72 direct CVS instructions (Schichow, Croker, Zimmerman, Höffler, & Härtig, 2016). Moreover, Lorch, Lorch, Freer, Calderhead, and Dunlap (2017) showed that CVS can be taught classroom-wide with such a short direct instruction, and that effects of such a training can even be maintained up to two-and-a-half years later. Their CVS-instruction used worked examples with good (i.e. valid) and bad (i.e. invalid or confounded) experimental designs (see Lorch et al., 2014 for the extensive description of methods). Both proved effective, but the condition where children evaluated bad experiment did maintain their learning gain better. Despite these studies that show the effectiveness of CVS instruction, it is still unknown whether the acquired experimentation abilities transfer to gains in new domains in the context of STEM education.

2. Training of teacher's verbal support

Another approach to strengthen children's scientific thinking is that teachers enrich children's reasoning processes by explicitly supporting their academic language use, for example, helping them to formulate their questions and by eliciting their arguments and ideas while reasoning. Over time, children can internalise such scaffolded reasoning, thus strengthening their own scientific thinking (Mercer, 2013). The rationale behind this can be found in the intertwined relation of science and language: Language allows for the creation of complex and abstract representations, and it helps children to represent what they see and do and subsequently allow them to discuss it (Mercer, 2013). Furthermore, science lessons are characterised by academic vocabulary and complex syntax structures, and children are required to use an extensive set of language tools like questioning, explaining, predicting and reasoning (e.g. Mercer, Dawes, Wegerif, & Sans, 2004). Indeed, several researchers demonstrated that children with strong linguistic abilities perform better in scientific tasks like CVS acquisition (Van der Graaf, Segers, & Verhoeven, 2016) and in transfer of the CVS strategy to new domains than children with weak linguistic abilities (Wagensveld et al., 2015).

Given the strong interrelatedness between children's linguistic and scientific abilities, several researchers advocate language-oriented science education (e.g. Barber et al., 2006). It has been argued that centralising language while teaching science may be beneficial for learning both science and language (Vitale & Romance, 2012). There are indeed strong indications that literacy and science education can be combined in the classroom to promote both of them (Clark & Lott, 2017). Different aspects of language have been integrated effectively in science education: reading (Vitale & Romance, 2012), writing (Nam, Choi, & Hand, 2011) and discourse (Mercer et al., 2004). A crucial component in all integrated instructions are teachers' verbal instruction strategies. It has been shown that teachers do not use these strategies by themselves in the context of science, due to a perceived lack of self-efficacy (Van Aalderen-Smeets, Walma van der Molen, & & Asma, 2012), but a teacher training enables in-service primary school teachers to apply verbal support strategies in the classroom (Smit, Gijsel, Hotze, & Bakker, 2018).

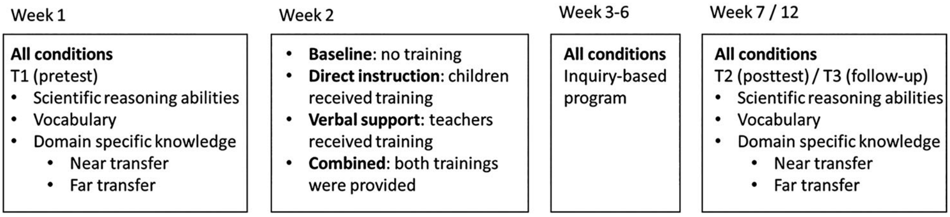
A crucial component in all integrated instructions are teachers' language strategies. In order to diagnose children's language and inquiry abilities and be responsive to it, teachers are required to actively use and promote the specialised science language, elicit children's explanations and provide feedback, and interactively discuss the inquiry process with students. These requirements are completely in line with the suggestions for enhanced instruction of inquiry learning that follow from a meta-analyses of effective inquiry learning (Alfieri et al., 2011). It has been shown that teachers can be effectively trained to apply these verbal scaffolding techniques and that, as a result, children's cooperation and understanding improves (Gillies, 2004). In addition, a teacher training on language promoting strategies might improve children's use of academic vocabulary (Henrichs & Leseman, 2014). However, it is yet unknown what children's learning gains with respect to the components of scientific reasoning and domain-specific knowledge is.

3. The present study

To summarise, inquiry-based lesson series are an ideal instructional method to teach children to reason scientifically and to let them acquire domain-specific knowledge in STEM (OECD, 2014). However, children have difficulties in applying their scientific reasoning abilities without guidance (Alfieri et al., 2011; Zimmerman, 2007) and as a consequence will have difficulties to further develop these abilities and to acquire knowledge within the STEM domains (Lazonder & Harmsen, 2016). Direct instruction beforehand as well as ongoing verbal support during children's inquiries can both be beneficial to develop and apply scientific thinking. It is easy to assume that combining the two may boost teaching of scientific thinking, but such combination has not before been investigated in one single intervention.

Therefore, in the present study we aimed to bridge the insights from research into direct instruction on scientific reasoning abilities with those from research into teacher training of verbal support. We investigated the combined effectiveness of both interventions for scientific knowledge and multiple scientific reasoning abilities, and compare this to the unique contribution of each and to a baseline (control) group. Two brief instructions were designed. The direct instruction consisted of an interactive instruction revolving around experiments and use of the CVS, but also addressing the hypothesis construction and evidence evaluation components. The teacher training consisted of instruction on the inquiry cycle and how verbal support can be provided in the context of inquiry-learning in STEM education. Teachers were instructed to implement linguistically rich contexts (e.g. emphasising academic vocabulary and language to structure the thought of the children), to stimulate children's elaborate responses, and to provide adequate feedback.

The effectiveness of both instructions was tested in a randomised controlled trial design with four groups: direct instruction only, verbal support only, combined condition, and a lesson-series-only (baseline) group (see Figure 1). Note that the baseline group thus is an inquiry-only group. As a follow-up, children were tested a few weeks after posttest. The



Week 1

All conditions
T1 (pretest)
- Scientific reasoning abilities
- Vocabulary
- Domain specific knowledge
  - Near transfer
  - Far transfer

Week 2

- **Baseline**: no training
- **Direct instruction**: children received training
- **Verbal support**: teachers received training
- **Combined**: both trainings were provided

Week 3-6

All conditions
Inquiry-based program

Week 7 / 12

All conditions
T2 (posttest) / T3 (follow-up)
- Scientific reasoning abilities
- Vocabulary
- Domain specific knowledge
  - Near transfer
  - Far transfer

Figure 1. Overview of the time frame and conditions.

following research questions were addressed: What are the effects of direct instruction, verbal support, and a combination of both on the effectiveness of an inquiry-based lesson series as assessed by children's a) scientific reasoning abilities, b) vocabulary, and c) domain-specific knowledge: near and far transfer? We expected the direct instruction group to improve more on scientific reasoning, mostly on the experimentation component of scientific reasoning (e.g. Chen & Klahr, 1999) and to a lesser extent on the other components. Additional improvement on domain-specific knowledge as acquired during the inquiry-based lesson series was also expected. Given the focus on the role of language in the verbal support condition, we expected larger gains on children's scientific reasoning and on their vocabulary (Henrichs & Leseman, 2014) compared to baseline condition. We expected the combined condition to show gains on all domains, as they are all addressed, and explored whether the combination would have a catalyst effect. On the long term, we also expected results to best consolidate in this group.

4. Methods

4.1. Participants

A convenience sample of schools in the Netherlands that were in the proximity to one of the research institutes agreed to participate. Together, there were 12 classrooms (10 schools) that took part in this study. This resulted in 301 participating children (164 girls and 137 boys), that were in 4th grade of primary school and who spoke Dutch. These children were 9 years and 5 months old on average (*SD* = 6 months) and generally came from middle to upper middle classrooms, as indicated by their parents' education (6.9% had attained basic education (ISCED level 0–2), 42.7% had attained upper secondary education (ISCED level 3–4), and 50.4% had attained tertiary education (ISCED level 5–8), in comparison to the national average, 26.6%, 41.3%, and 27.9%, respectively (Eurostat, 2017). To ensure a heterogeneous sample, all children in the classrooms participated in the study. Parents gave active consent for their child to participate.

4.2. Design

We conducted a randomised controlled trial design with a pretest (T1), direct posttest (T2) and long-term test (T3) (see Figure 1). The follow-up test was conducted five weeks after the inquiry-based lesson series. Classrooms were randomly assigned to one of the four conditions. These conditions differed in instruction before the inquiry-based lesson series: Child instruction (CI) or not and teacher training (TT) or not. All children received the inquiry-based lesson series in four successive weeks, twice a week for one hour with one week holiday in between. Children were assessed on scientific reasoning, vocabulary, and near and far transfer of domain-specific knowledge.

4.3. Materials

4.3.1. Intervention materials

4.3.1.1. Inquiry-based lesson series. The inquiry-based lesson series consisted of six inquiry-based science lessons (approximately 60 minutes each) on the topic of (solidity of) constructions. The inquiry-based lesson series was inspired by a similar themed lesson series as developed in a previous study (Gijsel, Vrielink, & Kiers, 2016). The main goal of the inquiry-based lesson series was threefold: to increase children's domain-specific knowledge concerning constructions, to foster their scientific abilities and to develop children's language abilities. To teach what determines the solidity of constructions, four topics (i.e. variables) were investigated during the lessons, using hands-on materials. These variables were: Materials, shapes, profiles, and connections.

To ensure that children could exercise with all scientific reasoning components, lessons were structured along the inquiry cycle. In the first lesson, children oriented on the topic of solid constructions and the design problem was introduced; building a solid bridge. The second and third lesson started with confrontation and exploration with the topic at hand (materials, shapes, profiles, and connections), followed by the generation of questions and hypotheses. For example, in the second lesson, children were introduced to profiles via a classroom discussion of what types of profiles exist and what kind of characteristics these have. Next, the teacher guided the discussion towards how profiles might be used to make constructions more solid. Children then started a short inquiry in small groups, in which they folded their own paper profiles (tube profile versus quadrilateral profile). They formulated a hypothesis concerning the solidity and set up an experimental design in which both types of profiles were compared concerning their capability to carry weight. Then, children conducted their experiments. Finally, children wrote down a conclusion followed by a classroom discussion. Worksheets guided children throughout all phases of the inquiry cycle. Afterwards, in the fourth lesson, children made a draft of their bridge, which was built, presented and discussed in lesson 5 and 6. Transfer was made to other constructions in the final lesson.

4.3.1.2. **Child instruction.** The aim of this direct instruction was to teach children how to recognise, design, and conduct valid experiments, via the learning of the Control of Variables Strategy (CVS). The instruction was based on the invalid experiments instruction by Lorch and colleagues (2014) and adapted to suit the age of children and the goals of this study. The instructor (i.e. the first author of this paper) was introduced to the children as a guest teacher for one day. He first introduced the topic, which was experimentation. Next the materials were introduced, which were two large, wooden ramps with four variables each (see Chen & Klahr, 1999). This introduction was done by setting up an invalid experiment, where not all variables were controlled, so children could identify which variables were manipulated and which were controlled. The next step was to correct the experimental set-up, which was done interactively. The example was concluded by letting the balls roll down the slope of the ramps.

The main component of the instruction was correcting two invalid experiments. After introduction of the invalid experiment, children recreated the settings of one of the ramps using their own small replica of the wooden ramps. This was done in pairs. Children were asked to set up the other ramp correctly for a valid experiment. They used worksheets for this. The instructor and the teacher aided in this process by drawing children's attention to what a correct design is and what the settings of the variables were. Each experiment ended with letting the balls roll and drawing conclusions about the variable that was investigated. This was done by the children and thereafter by the instructor using the two large, wooden ramps in front of the class. We chose for this activity to draw attention to the role of experiments in the inquiry process, because by conducting the experiment and drawing conclusions based on the results one can decide whether the research question was answered. Another reason was to motivate the children.

A third experiment was included to help children to transfer the CVS strategy from the wooden ramps experiment to another domain. For this a paper-and-pencil test was used with an experiment about drink sales (Chen & Klahr, 1999). Children could choose how they would set up two lemonade stalls, two possible opening times of the stand (at 12:00 or at 15:00), whom to sell the lemonade to (older or younger children), and what type of drink they would sell (lemonade or iced tea). Children were asked to design an experiment that investigated the effect of opening hours on how many drinks would be sold. After they filled in their answers on the worksheet, the experiment was discussed interactively. Finally, it was concluded that experiments can be invalid and they can be made valid by applying the CVS. Also, the experimenter stressed that experimentation strategies like during the experiments with the wooden ramps can be used in other settings, such as drink sales and many more.

4.3.1.3. **Teacher training.** The teacher training on verbal support was a 3-hour training, led by the third author of this paper, who drew on her experience in language-oriented science lessons. The training was designed around four types of learning experiences: experience, reflection, conceptualising, and application (Bijkerk & Van der Heide, 2006). The main goal of the training was to increase teachers' awareness of the role of language in science and to provide them with instructional strategies to promote the quality of children's language during their science lessons. First, the nature of science was elucidated and teachers got acquainted with the basic elements of inquiry-based science education and the major role of language in science. A distinction was made between school vocabulary, domain-specific vocabulary, and research-related vocabulary. Secondly, participants watched a video in which another teacher provided a language-oriented science lesson about sound and acoustics to Grade 6 children. We selected this video, because it included several good examples of instructional strategies. Participants reflected on these strategies to promote children's language. Afterwards, a group discussion about teachers' observations took place. Finally, three components of language-developing strategies were explained in detail and illustrated with short video's (Verhallen & Walst, 2001): (1) teachers' use of language and questioning; (2) teachers' interactional abilities; and (3) teachers' feedback. At the end of the training, all teachers were provided with hand-outs of the presentation.

4.3.2. Measures

4.3.2.1. **Scientific reasoning abilities.** The Scientific Reasoning Inventory (SRI) measures components of scientific reasoning in a paper-and-pencil multiple choice task. The task was based on the Argumentative Sensemaking Measure developed for middle school (Bathgate, Crowell, Schunn, Cannady, & Dorph, 2015) and adapted to elementary school by Van de Sande, Verhoeven, Kleemans, and Segers (in press). The subtasks are based on validated scientific reasoning tasks from previous research (Chen & Klahr, 1999; Kuhn & Dean, 2005; Schröder, Bödeker, & Edelstein, 2000). The test was shown to be reliable and valid in 11- to 12-year-olds, i.e. Grade 6 (Van de Sande et al., in press). The first subtask measured coordination of theory and evidence (hypothesis generation and evidence evaluation). In this subtask children received a text and had to select one of four alternative research questions that best covers the research in the story. Other type of questions were to interpret data from a matrix. The second subtask was an assessment of how skilled children were in applying the Control of Variables Strategy. The final subtask was a syllogistic reasoning task.

One item was added in the present study to complement a story with one question about syllogistic reasoning, so all stories had two questions. The SRI in the present study consisted of 9 items about coordination of theory and evidence, 7 items about experimentation, and 8 items about syllogistic reasoning. The syllogistic reasoning items were presented in pairs. Each pair was about the same proposition.

4.3.2.1.1. Confirmatory factor analysis. To test whether the inventory covered the same components in Grade 4 in the present study as in the Grade 6 of the previous study (Van de Sande et al., in press), a confirmatory factor analysis (CFA) was performed. A model was built with three components and the corresponding items. The components were allowed to covary. Covariances were added between syllogistic reasoning items that were pairs. Four items did not load as expected and they were removed from the model. Those were two syllogistic reasoning items and two experimentation items. The final model had a strong fit (Tabachnick & Fidell, 2001), $\chi^2(164) = 153.14$, $p, = .718$, CFI = 1.00, TFI = 1.02, RMSEA < .001, SRMR = .050. Only one question did not load significantly on the component of experimentation, $p = .058$. As this was close to significance and in the expected direction, the path was not omitted from the model. Therefore, it can be concluded that the same components of the SRI were found in Grade 4 as in Grade 6, which confirms the SRI's content validity.

Reliability was assessed for the SRI as a whole and per component (see Table 1). All coefficients indicated good reliability, except for experimentation at T1, but reliability was good at T2 and T3.

4.3.2.2. Vocabulary.

This test was developed to measure the academic and domain-specific vocabulary that was taught and used during the inquiry-based lesson series about constructions and consisted of 25 questions about academic and domain-specific vocabulary, such as what is a research question and what is an architect. All questions were multiple-choice questions with four options. The maximum score was 25 correct. Reliability was good, Cronbach's $\alpha = .60$ (T1), .75 (T2), and .77 (T3), Guttman's $\lambda_2 = .61$ (T1), .77 (T2), and .78 (T3).

4.3.2.3. Domain-specific knowledge: near transfer.

Near transfer was assessed through the domain-specific knowledge that children gained with the inquiry-based lesson series about constructions. Two tests were developed for the current study, the tower test and

Table 1. Reliability coefficients of the SRI and its components.

	Cronbach's α			Guttman's λ_2		
	T1	T2	T3	T1	T2	T3
Scientific reasoning inventory ($n = 20$)	.64	.83	.84	.67	.84	.85
Coordination of evidence and theory ($n = 9$)	.63	.67	.71	.64	.68	.71
Experimentation ($n = 5$)	.39	.74	.74	.40	.74	.74
Syllogistic reasoning ($n = 6$)	.80	.85	.89	.81	.85	.89

the bridge test. The tower test aimed at assessing children's active knowledge about constructions. Through an open-ended question they were asked to indicate as much as possible about how they would build a tower that was tall and strong. To score this assessment, we developed a strict annotation protocol, in collaboration with a professional construction engineer, that followed the dimensions as included in the inquiry-based lesson series (i.e. materials, shapes, profiles, and connections) and four other main categories of constructions. For each of these variables that the children mentioned they received one point. Two independent raters rated the answers of approximately 10% of the subjects. Unsure cases were discussed with the authors. The overall agreement of the raters was moderate to substantial (Landis & Koch, 1977), the intraclass correlation (ICC) was .72, and Cohen's $\kappa = .55$. A maximum of eight points could be scored on the tower test.

The second test of near transfer, the bridge test, was a set of six multiple-choice questions. Children were asked to identify the strongest bridge out of four bridges. Within these options we varied the dimensions that were taught in the inquiry-based lesson series. One item was removed, because it showed an item-total correlation that was too low. The following reliability analysis showed acceptable reliability, Cronbach's $\alpha = .19$ (T1), .66 (T2), and .66 (T3), Guttman's $\lambda_2 = .30$ (T1), .68 (T2), and .68 (T3), except for T1, but this can be explained by the complete novelty of the required knowledge. The maximum number of correct was five.

4.3.2.4. Domain-specific knowledge: far transfer.

Far transfer was assessed with a test about constructions, but with a topic that was not part of the inquiry-based lesson series. Four different objects were presented and the task was to explain how these could be made stronger. Just as for the tower test, the answer were scored with an answer sheet. The four variables that were part of the inquiry-based lesson series were scored. Reliability analyses revealed that there was low agreement between the raters on one of the four objects. The low agreement remained after one additional instruction by one of the researcher. Therefore, the object was removed from analyses. The three remaining objects showed fair to substantial interrater reliability (Landis & Koch, 1977), the ICC was .64 and Cohen's $\kappa = .32$. The total score was 12 points.

4.4. Procedure

We conducted a small pilot of the inquiry-based lesson series before the start of this study, which led to some small, additional improvements. Testing took place in the classroom and children filled in paper-and-pencil tests. This was done in two sessions of approximately 45 minutes with a break in between. One of the researchers or research assistants supervised these sessions together with the teacher of that class. Before the lesson series and after the pretest, there was child-based instruction, teacher training, a combination of both, or neither of them, depending on the condition the classroom was in. The child instruction was an one hour lesson that was provided in the classroom by the first author. The teacher training took place at one of the research institutes. Teachers visited the research institute for a three hour training session provided by the third author. The inquiry-based lesson series consisted of six lessons (approximately 60 minutes each) in a period of four weeks and took place between the pretest and posttest. In all conditions, only the teacher guided the inquiry-based lesson series. Multiple actions were undertaken to check the treatment fidelity. First, a random selection of the classrooms (eight lessons in total) was visited by two researchers to observe the lessons. These lessons were audio- and videotaped. During the observations, the researcher was seated out of sight of the children as much as possible. Secondly, teachers were asked to keep track of their lessons and their experiences in a logbook. Third, teachers were interviewed about their experiences with the lesson series afterwards. Topics in the interviews included employability of the instruction, teachers' preparations and implementation of the lesson series (amount of time, use of additional materials), an evaluation of the lesson goals and an overall evaluation of the lesson series. In all, these three information sources indicated that the teachers followed the instructions about the lesson series provided by the researchers and that classroom processes were comparable. It was unlikely that the teachers spent additional time on either the topic (solid constructions) or on inquiry learning, as it is more time-consuming than their usual science lessons. The posttest took place in the week following the inquiry-based lesson series, while the long-term posttest was assessed 5 weeks after the posttest.

5. Results

5.1. Effects of instructional condition on the learning gains

Before analyses were conducted, missing values in the data were inspected. Visual inspection indicated that the missing data was mostly grouped per measurement, indicating that the participant was absent during one specific measurement. There were no further indications of patterns in the missing data. The analyses performed in this study excluded

Table 2. Means and standard deviations (between brackets) for scientific reasoning, vocabulary, and near and far transfer of domain-specific knowledge over time and per condition.

		Baseline	Direct instruction	Verbal support	Combined
		M (SD)	*M (SD)*	*M (SD)*	*M (SD)*
Scientific reasoning					
Coordination of theory and evidence	T1	4.49 (2.13)	4.63 (2.34)	4.95 (2.28)	4.19 (2.10)
	T2	4.82 (2.34)	5.16 (2.52)	5.87 (1.92)	4.83 (2.42)
	T3	5.11 (2.31)	5.26 (2.43)	5.99 (2.10)	4.53 (2.37)
Experimentation	T1	1.30 (1.20)	1.25 (1.14)	1.27 (1.13)	1.29 (1.17)
	T2	1.36 (1.31)	3.11 (1.73)	2.20 (1.52)	3.07 (1.66)
	T3	1.79 (1.54)	2.92 (1.71)	2.27 (1.71)	2.79 (1.74)
Syllogistic reasoning	T1	3.41 (2.12)	3.68 (2.07)	4.20 (1.94)	3.44 (1.99)
	T2	4.24 (1.96)	4.15 (2.15)	4.61 (1.99)	4.06 (2.09)
	T3	4.65 (2.13)	4.25 (2.20)	4.95 (1.83)	4.11 (2.18)
Vocabulary	T1	11.12 (3.90)	11.37 (3.29)	12.94 (3.70)	12.12 (3.09)
	T2	13.54 (4.47)	13.53 (4.71)	15.20 (4.39)	15.32 (3.38)
	T3	13.83 (4.38)	14.40 (4.74)	15.30 (4.73)	14.50 (4.03)
Near transfer					
Bridge	T1	0.56 (0.77)	0.83 (0.88)	0.83 (1.00)	0.69 (0.75)
	T2	1.81 (1.61)	1.86 (1.38)	1.25 (1.27)	2.01 (1.56)
	T3	1.73 (1.52)	1.51 (1.45)	1.21 (1.39)	1.32 (1.41)
Tower	T1	1.39 (0.94)	1.13 (0.89)	1.37 (0.95)	1.43 (0.89)
	T2	2.29 (1.11)	2.21 (1.21)	2.65 (0.94)	2.49 (1.25)
	T3	2.03 (1.01)	2.00 (0.96)	2.19 (0.94)	2.49 (1.12)
Far transfer					
Objects	T1	2.68 (1.08)	2.22 (1.14)	2.49 (1.03)	2.33 (1.16)
	T2	2.87 (1.43)	2.31 (1.28)	2.59 (1.35)	2.52 (1.44)
	T3	2.75 (1.40)	2.41 (1.34)	2.45 (1.17)	3.00 (1.34)

missing values. Table 2 gives an overview of the descriptive statistics of the dependent variables over time, across the four conditions.

To examine learning gains in scientific thinking (i.e. scientific knowledge and scientific reasoning), paired samples t-tests were conducted to reveal overall progress over time (from T1 to T2). Next, it was investigated whether the conditions had an effect on the change over time by conducting regression analyses with planned contrasts on performance at T2. To control for possible differences between children at T1, the first predictor in the regression analysis was performance on each specific test at T1. Planned contrasts were used to investigate effects of direct instruction and verbal support compared to baseline, as well as the combined condition. The planned contrasts were direct instruction versus baseline, verbal support versus baseline, and combined versus baseline. These contrasts were entered as additional predictors into the regression analyses. Conditions with instruction (direct instruction, verbal support, or combined) were dummy coded as 1 and baseline as −1, while the conditions that were not relevant for the specific contrast were coded as 0. This way, positive regression coefficients indicated that the condition with instruction performed better than the baseline condition.

Regarding *scientific reasoning*, performance on T2 was higher than on T1 for all components of scientific reasoning (coordination of theory and evidence, $t(267) = 4.84$, $p < .001$, $d = 0.28$, experimentation, $t(266) = 8.78$, $p < .001$, $d = 0.78$, and syllogistic reasoning, $t(264) = 4.45$, $p < .001$, $d = 0.27$). The components of scientific reasoning at T2 were predicted by their performance on T1. Coordination of theory and evidence was also predicted by the contrast verbal support versus baseline and experimentation by the contrast direct instruction versus baseline, whereas neither condition related to syllogistic reasoning at T2 (see Table 3). The positive values of the contrasts indicated that the instructions (direct instruction and verbal support) contributed positively to performance at T2 compared to baseline.

Regarding *vocabulary*, performance on T2 was better than on T1, $t(269) = 13.17$, $p < .001$, $d = 0.66$. T1 predicted T2, as well as the contrast of combined versus baseline (see Table 3), indicating that the combination of instructions was more effective than baseline.

Near transfer performance on T2 was better than on T1: bridge test, $t(267) = 10.37$, $p < .001$, $d = 0.89$, and tower test, $t(266) = 13.23$, $p < .001$, $d = 1.01$. Near transfer on T2 was predicted by its performance on T1, the contrast between combined and baseline, and the contrast between verbal support and baseline. The beta-value of the contrast between verbal support and baseline suggested that the baseline condition improved more on domain-specific knowledge than verbal support, because the coefficient was negative (see Table 3). However, given that the means differ only slightly (see Table 2), and the possibility of suppression effects in regression analysis, we performed an extra check. A regression analysis with T1 and only the contrast of verbal support versus baseline as predictors, showed that the contrast did not significantly predict performance on domain-specific knowledge at T2, $p = .796$. The tower test on T2 was predicted by T1, but not by the planned contrasts.

Finally, regarding *far transfer*, the children scored better on T2 than on T1, $t(284) = 2.01$, $p = .037$, $d = 0.14$. Performance on T2 was predicted by performance on T1, and not by the planned contrasts.

Table 3. Dependent variables (in cursive), independent variables, and their B and ⍰-values for the regression analyses of T2, as well as the amount of explained variance of the regression.

Independent variable	R^2	*B*	SE (*B*)	β
Coordination of theory and evidence T2				
Coordination of theory and evidence T1		0.56	.05	.53**
Direct instruction versus baseline		−0.14	.21	−.04
Verbal support versus baseline		0.57	.21	.17**
Combined versus baseline		−0.08	.21	−.02
	.31			
Experimentation T2				
Experimentation T1		−0.07	.08	−.05
Direct instruction versus baseline		0.65	.16	.27**
Verbal support versus baseline		−0.18	.17	−.07
Combined versus baseline		0.60	.16	.25**
	.17			
Syllogistic reasoning T2				
Syllogistic reasoning T1		0.51	.05	.51**
Direct instruction versus baseline		−0.16	.19	−.06
Verbal support versus baseline		0.11	.20	.04
Combined versus baseline		−0.06	.19	−.02
	.27			
Vocabulary T2				
Vocabulary T1		0.82	.06	.67**
Direct instruction versus baseline		−0.43	.33	−.07
Verbal support versus baseline		−0.10	.35	−.02
Combined versus baseline		0.84	.33	.14*
	.47			
Domain-specific knowledge: near transfer				
Bridge test T2				
Bridge test T1		0.29	.10	.17**
Direct instruction versus baseline		0.09	.15	.04
Verbal support versus baseline		−0.47	.16	−.22**
Combined versus baseline		0.35	.15	.17**
	.07			
Tower test T2				
Tower test T1		0.26	.08	.21**
Direct instruction versus baseline		−0.17	.12	−.11
Verbal support versus baseline		0.23	.12	.14
Combined versus baseline		0.07	.12	.04
	.07			
Domain-specific knowledge: far transfer				
Objects test T2				
Objects test T1		0.39	.07	.32**
Direct instruction versus baseline		−0.22	.13	−.11
Verbal support versus baseline		−0.04	.13	−.02
Combined versus baseline		0.01	.13	.01
	.13			

**p* < .05, ***p* < .01.

5.2. Follow-up test

As a follow-up, the children were tested six weeks after the inquiry-based lesson series and thus five weeks after posttest (T2). The same analysis plan as for T2 was followed, which means that paired samples t-tests and regression analyses were performed.

Performance on T3 was better than on T1 for all components of scientific reasoning (coordination of theory and evidence, $t(267) = 5.32$, $p < .001$, $d = 0.30$, experimentation, $t(265) = 9.38$, $p < .001$, $d = 0.84$, and syllogistic reasoning, $t(261) = 6.09$, $p < .001$, $d = 0.36$). Performance on T3 was the same as on T2 for coordination of theory and evidence, $t(261) = 0.96$, $p = .339$, $d = 0.05$, and experimentation, $t(261) = 0.39$, $p = .698$, $d = 0.02$.

Performance did increase slightly on the syllogistic reasoning component, $t(257) = 2.45$, $p = .015$, $d = 0.12$. The same effects were found for T3 as for T2: T1 predicted performance at T3 for all components of scientific reasoning (see Table 4). Verbal support versus baseline predicted *coordination of theory and evidence*. *Experimentation* was predicted by direct instruction versus baseline and combined versus baseline. Whether the performance at T3 was better in the direct instruction or combined condition, while taking T1 into account, was investigated by an additional regression analysis. The planned contrast combined versus direct instruction and experimentation at T1 were the predictors. The effects were not significant, which is in line with the comparable increase from T1 to T3 in both conditions (see Table 2). No planned contrast predicted *syllogistic reasoning*.

Table 4. Dependent variables (in cursive), independent variables, and their B and ⍰-values for the regression analyses of T3, as well as the amount of explained variance of the regression.

Independent variable	R^2	B	SE (B)	⍰
Coordination of theory and evidence T3				
Coordination of theory and evidence T1		0.60	.05	.57**
Direct instruction versus baseline		−0.07	.20	−.02
Verbal support versus baseline		0.61	.20	.18**
Combined versus baseline		−0.41	.20	−.12
	.37			
Experimentation T3				
Experimentation T1		−0.14	.09	−.10
Direct instruction versus baseline		.41	.17	.17*
Verbal support versus baseline		−0.09	.18	−.03
Combined versus baseline		0.35	.17	.15*
	.07			
Syllogistic reasoning T3				
Syllogistic reasoning T1		0.55	.05	.53**
Direct instruction versus baseline		−0.26	.19	−.09
Verbal support versus baseline		0.24	.19	.08
Combined versus baseline		−0.18	.19	−.06
	.30			
Vocabulary T3				
Vocabulary T1		0.74	.07	.57**
Direct instruction versus baseline		0.35	.38	.06
Verbal support versus baseline		0.19	.40	.03
Combined versus baseline		−0.06	.38	−.01
	.34			
Domain-specific knowledge: near transfer				
Bridge test T3				
Bridge test T1		0.38	.10	.23**
Direct instruction versus baseline		0.00	.15	.00
Verbal support versus baseline		−0.25	.15	−.12
Combined versus baseline		−0.09	.15	−.04
	.07			
Tower test T3				
Tower test T1		0.11	.07	.10
Direct instruction versus baseline		−0.20	.11	−.15
Verbal support versus baseline		0.05	.11	.04
Combined versus baseline		0.33	.12	.22**
	.05			
Domain-specific knowledge: far transfer				
Objects test T3				
Objects test T1		0.35	.07	.29**
Direct instruction versus baseline		−0.24	.13	−.13
Verbal support versus baseline		−0.23	.13	−.13
Combined versus baseline		0.44	.15	.22
	.12			

*p < .05, **p < .01.

Scores on the *vocabulary* test at T3 were higher than T1, $t(268) = 11.14$, $p < .001$, $d = 0.64$, and did not differ from T2, $t(261) = 0.11$, $p = .915$, $d < 0.01$. Performance on the vocabulary test at T3 was only predicted by its performance on T1. Scores on *near transfer of domain-specific knowledge* were higher at T3 than T1: bridge test, $t(265) = 7.60$, $p < .001$, $d = 0.59$, and tower test, $t(241) = 10.87$, $p < .001$, $d = 0.93$, but lower at T3 compared to T2: bridge test, $t(260) = 3.12$, $p = .002$, $d = 0.21$, and tower test, $t(236) = 2.21$, $p = .028$, $d = 0.17$. Near transfer, as measured with the bridge test, at T3 was only predicted by itself at T1. The tower test was predicted by the contrast of combined versus baseline. Scores on the *far transfer test* were higher at T3 than T1, $t(259) = 2.93$, $p = .004$, $d = 0.22$, and did not differ from T2, $t(273) = 0.92$, $p = .358$, $d = 0.06$. Far transfer at T3 was significantly predicted by both T1 and the contrast between combined and baseline.

6. Discussion

The aim of the present study was to investigate the effects of a short child-based direct instruction on scientific reasoning and a training of teacher's verbal support, combined or in isolation, as pre-instructions to the effects on scientific thinking of a STEM-based lesson series. The effectiveness was evaluated by assessing children's scientific thinking, i.e. their scientific reasoning abilities and scientific knowledge (vocabulary and domain-specific knowledge). The results revealed that both the direct instruction on scientific reasoning and ongoing verbal support during the full process of inquiry learning boosted the effectiveness of the inquiry-based lesson series and that, as expected, the combined condition led to the largest overall learning gains.

6.1. Scientific reasoning abilities

As expected, additional and in some cases differential effects of both instruction were found to the different scientific reasoning abilities.

6.1.1. Coordination of theory and evidence

With reference to coordination of theory and evidence, all groups increased over time, but direct instruction and the combined condition did not show an additional effect compared to the baseline condition, while the verbal support condition did on both posttest and follow-up.

Contrary to our hypothesis, the direct instruction did not result in additional improvements on the coordination of theory and evidence. This may be due to the fact that the direct instruction was mainly focused on experimentation, while other aspects of scientific reasoning were addressed more implicitly. As a result, the direct instruction affected experimentation (in line with Lorch et al., 2014), but did not transfer to an increase in the coordination of theory and evidence. This explanation should be interpreted with caution, because the way these abilities are assessed appears to be related to the effectiveness of instruction. In the present study, the assessment did not incorporate the link between hypotheses and evidence on one hand with experimentation on the other hand. When this link is more explicitly part of the assessment, effects of direct instruction can be found (Klahr & Nigam, 2004).

As expected, the verbal support condition did have an additional effect compared to baseline. Others have also found that verbal support can help children in their scientific activities (Clark & Lott, 2017), but the present study is the first to show that stronger verbal support leads to an additional improvement on coordination of theory and evidence. It should be noted that teachers had to make a transfer themselves after the training to inquiry in the classroom. The effect of verbal support on coordination of theory and evidence is in line with other studies showing the dependency of scientific reasoning on language and linguistic abilities, such as the critical role of meaning in deductive reasoning (Kuhn & Franklin, 2007), and the role of the teacher herein (Mercer et al., 2004). Another argument for the effectiveness of verbal support on the coordination of theory and evidence is that the inquiry cycle starts with the hypothesis, where this new process of inquiry is introduced, as well as the new concepts about which hypothesis should be constructed and finally outcomes should be translated into theoretical frameworks, which are verbal representations. Children describe the world around them using models that are more than descriptions of the observable and also include unseen hypothetical entities that interact to produce emergent behaviour (Lehrer & Schauble, 2000). Language can support such models, as it enables abstract and complex forms of thought (Mercer, 2013).

It is surprising, therefore, that the combined condition did not show an additional improvement on coordination of theory and evidence. Perhaps the inclusion of the direct instruction focused both teachers (who were also present during that instruction) and children on experimentation, which might make them talk more about experimentation and therefore less about coordination of theory and evidence. This speculation calls for further examination through classroom observations to get a better understanding of the ongoing process.

6.1.2. Experimentation

With regard to experimentation, the direct instruction and combined condition showed the largest learning gains on posttest and follow-up. The verbal support did not differ from baseline.

Children that received the direct instruction either in isolation of in the combined condition improved more than the baseline condition, and the improvement did retain in the long-term compared to baseline. This result is in line with previous studies that show effectiveness of only a brief CVS intervention to the development of experimentation abilities in the short-term (Chen & Klahr, 1999; Wagensveld et al., 2015) and long-term (Lorch et al., 2017). This result can be explained by the relevancy of direct instruction in experimentation (Schichow et al., 2016). The present results add to the existing literature that such an instruction improves the effectiveness of an inquiry-based lesson series in teaching children experimentation abilities compared to no preceding instruction before the inquiry-based lesson series.

There were no effects of verbal support on experimentation. The verbal support aimed to help children explain and reason during their inquiries, but it might not have remedied children’s natural tendencies to manipulate more variables than one (Wilkening & Huber, 2004). This does not mean that experimentation does not depend on verbal factors (e.g. Van der Graaf et al., 2016; Wagensveld et al., 2015), only that experimentation is an ability that requires a direct focus.

6.1.3. Syllogistic reasoning

Although syllogistic reasoning developed over time in all children, there were no differential effects per condition. For additional improvement on syllogistic reasoning, a different instruction may be needed. The verbal approach seemed promising, as brain activation during syllogistic reasoning revealed a verbal network, but when semantic content is lacking, a different network is recruited that suggests that abstract syllogisms are solved using spatial models (Goel, Buchel, Firth, & Dolan, 2000). Syllogistic reasoning, therefore, may be more a part of general ability than the other factors in the construct of scientific reasoning.

6.2. Vocabulary

Only the combined direct instruction and verbal approach showed effective to further strengthen vocabulary gains. The findings that either instruction alone did not lead to an increase in vocabulary might be explained by their synergetic effects in that both have different strengths, but those do no lead to a significant improvement. While direct instruction on scientific reasoning did not explicitly focus on vocabulary, it might have strengthened vocabulary learning by providing children the reasons to learn and use new concepts. Verbal support, on the other hand, did focus on vocabulary use, but teachers were not instructed to highlight why the concepts were relevant. Therefore, the combination showed an improvement in vocabulary. Direct instruction could have helped in understanding why concepts should be learning. It has been shown that highlighting the ‘why’, i.e. promoting the purpose for learning, improved academic achievement in science and mathematics (Yeager et al., 2014). Verbal support helped in learning and using the concepts by eliciting explanations and providing contexts. Both explanation and context stimulate vocabulary learning of primary school children (Cain, 2007).

Vocabulary at the follow-up was not predicted by any contrast. The scores were higher than on the pretest. It might be that the words in the vocabulary test are not usually used in the classroom and that our inquiry-based lesson series was an exception. Most words were related to the topic, i.e. solid construction, or to inquiry. Since the topic was dealt with, it is likely that classes moved forward to a new topic. Also, inquiry learning is not common in primary school classrooms (OECD, 2014).

6.3. Domain-specific knowledge: near and far transfer

In line with the results for vocabulary, only domain-specific knowledge effects were found when combining both types of instruction. Direct instruction might have helped in setting up and understanding experiments and verbal support in guiding the inquiry, including drawing of conclusions. This results is in line with findings that the inquiry process can be stimulated leading to knowledge acquisition (Lazonder & Harmsen, 2016), and explains why the combination showed improvements in domain-specific knowledge on both near and far transfer while the individual instructions did not.

For the posttest effects were found for one test of near transfer and on the follow-up for the other test of near transfer. An explanation for this result might be the type of questions that were used. The first test used was a multiple choice task, while the second one consisted of open-ended questions. Given that open-ended questions measure the quality of active processing of relevant concepts during learning (Ozuru, Briner, Kurby, & McNamara, 2013), it can be assumed that the combined condition promoted active processing leading to more robust knowledge. This effect is visible in the follow-up, because some knowledge seems to be forgotten in the other conditions, but not in the combined condition. On the other hand, multiple choice relies more on topic-specific knowledge (Ozuru et al., 2013). This would be present directly after learning and therefore the combined condition showed an effect on the multiple choice questions at posttest.

What is left to be explained is how the near transfer effects moved from one test (the bridge test) at posttest to the other test (the tower test) at follow-up. These tests differed in the response options, the bridge test consisted of multiple choice questions and the tower test was an open-ended question. Multiple choice questions rely more on topic-specific knowledge and open-ended questions measure the quality of active processing of relevant concepts during learning (Ozuru et al., 2013). Topic-specific knowledge would be present directly after learning and effects on learning as well. In contrast, active processing would lead to more robust knowledge and effects of improved active processing in the combined condition are more easily visible at a follow-up test, because there is no forgetting, while children in the other conditions appear to forget some knowledge.

Interestingly, the combined condition led to improvements on the near, but not the far transfer task on posttest, and vice versa for the follow-up. This indicated that time had an effect on how the knowledge was stored; from episodic to long-term memory. Memories can undergo changes over time, such as an emergence of awareness for what had been learned earlier (Robertson, 2009). In addition, performance has been shown to be affected by metacognitive awareness via strategic control (Whitebread, 1999). Metacognitive awareness was likely to be boosted by verbal support, as it promoted reflection. Strategic control was improved by the direct instruction on scientific reasoning.

6.5. Limitations and suggestions

The present study is the first to show that pre-instructions improve the effectiveness of an inquiry-based lesson series as revealed by higher scores on coordination of theory and evidence, experimentation, vocabulary, and domain-specific knowledge: near and far transfer. However, some limitations can be identified. One is that we did not study the effectiveness of the inquiry-based lesson series compared to a business as usual control group. The present results showed an overall improvement, which indicated that the inquiry-based lesson series was effective in stimulating children's scientific thinking. Although unguided inquiry often is not effective (Kirschner et al., 2006), a comparison would allow for conclusions about the inquiry-based lesson series. This is a topic for future research. Another limitation is that the quality of the inquiries during the lesson series was not formally assessed, nor the role of the teachers. Video-tapes of lessons would provide rich information, which would add to the understanding of the underlying processes. It would be interesting to study to whether the quality of the conclusions predicts domain-specific knowledge (Van der Graaf, Segers, & Verhoeven, 2018). Finally, due to practical reasons the direct instruction on scientific reasoning was provided by one of the authors. It therefore remains to be investigated whether teachers can apply such instruction themselves to the children in their own classroom with brief instructions, or whether they need further training beforehand.

6.6. Conclusion and implications

To conclude, inquiry learning in primary education is feasible and both a direct instruction of scientific reasoning and verbal support during inquiry learning improved the effectiveness of an inquiry-based lesson series. Moreover, the combination of both strengthened the effects on scientific reasoning and scientific knowledge most. Children showed larger learning gain in scientific reasoning, vocabulary, and domain-specific knowledge. Therefore, we suggest using inquiry learning in the primary school classroom and prepare it by instructing children about experimentation and training teachers in verbal support. This way, children can train important twenty-first-century skills, such as scientific reasoning, and independently acquire scientific knowledge.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This research was funded by a grant from NRO [grant number 405-16-514].

References

Alfieri, L., Brooks, P. J., Aldrich, N. J., & Tenenbaum, H. R. (2011). Does discovery-based instruction enhance learning? *Journal of Educational Psychology*, *103*(1), 1–18. doi:10.1037/a0021017

Andersen, C., & Garcia-Mila, M. (2017). Scientific reasoning during inquiry. In K. S. Taber & B. Alpan (Eds.), *Science education. new directions in mathematics and science education* (pp. 105–117). Rotterdam: Sense. doi:10.1007/978-94-6300-749-8_8

Barber, J., Catz, K. N., & Arya, D. (2006, April). *Improving science content acquisition through a combined science/literacy approach: A quasi-experimental study*. Paper presented at the annual meeting of the American Educational Research Association, San Fransisco, CA.

Bathgate, M., Crowell, A., Schunn, C., Cannady, M., & Dorph, R. (2015). The learning benefits of being willing and able to engage in scientific argumentation. *International Journal of Science Education*, *37*(10), 1590–1612. doi:10.1080/09500693.2015.1045958

Bijkerk, L., & Van der Heide, W. (2006). *Het gaat steeds beter. Activerende werkvormen voor de opleidingspraktijk [It is getting better. Activating teaching methods for educational practice]*. Houten: Bohn Stafleu van Loghum.

Cain, K. (2007). Deriving word meanings from context: Does explanation facilitate contextual analysis? *Journal of Research in Reading*, *30*(4), 347–359. doi:10.1111/j.1467-9817.2007.00336.x

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. doi:10.1111/1467-8624.00081

Clark, S. K., & Lott, K. (2017). Integrating science inquiry and literacy instruction for young children. *The Reading Teacher*, *70*(6), 701–710. doi:10.1002/trtr.1572

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

Edelsbrunner, P., Schalk, L., Schumacher, R., & Stern, E. (2015). *Pathways of conceptual change: Investigating the influence of experimentation skills on conceptual knowledge development in early science education*. Pasadena, CA: Cognitive Science Society.

Eurostat. (2017). *Population by educational attainment level, sex and age* [Dataset]. Retrieved from http://ec.europa.eu/eurostat/product?code = edat_lfs_9903&language = en&mode = view

Fischer, F., Kollar, I., Ufer, S., Sodian, B., Hussmann, H., Pekrun, R., … Eberle, J. (2014). Scientific reasoning and argumentation: Advancing an interdisciplinary research agenda in education. *Frontline Learning Research*, *2*(3), 28–45. doi:10.14786/flr.v2i3.96

Gijsel, M., Vrielink, F., & Kiers, L. (2016). Taalgericht w&t-onderwijs in de onderbouw van het SBO [Language-oriented S&T education in the lower grades of special education]. In S. Koenen (Ed.), *Taal in de context van w&t [Language in the context of S&T]* (pp. 43–47). The Hague: Platform Bèta Techniek.

Gillies, R. M. (2004). The effects of communication training on teachers' and students' verbal behaviours during cooperative learning. *International Journal of Educational Research*, *41*(3), 257–279. doi:10.1016/j.ijer.2005.07.004

Goel, V., Buchel, C., Firth, C., & Dolan, R. J. (2000). Dissociation of mechanisms underlying syllogistic reasoning. *NeuroImage*, *12*(5), 504–514. doi:10.1006/nimg.2000.0636

Henrichs, L. F., & Leseman, P. P. (2014). Early science instruction and academic language development can go hand in hand. The promising effects of a low-intensity teacher-focused intervention. *International Journal of Science Education*, *36*(17), 2978–2995. doi:10.1080/09500693.2014.948944

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist*, *41*(2), 75–86. doi:10.1207/s15326985ep4102_1

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

Klahr, D., & Nigam, M. (2004). The equivalence of learning paths in early science instruction: Effects of direct instruction and discovery learning. *Psychological Science*, *15*(10), 661–667. doi:10.1111/j.0956-7976.2004.00737.x

Klahr, D., Zimmerman, C., & Jirout, J. (2011). Educational interventions to advance children's scientific thinking. *Science*, *333*(6045), 971–975. doi:10.1126/science.1204528

Kuhn, D., & Dean, D. (2005). Is developing scientific thinking all about learning to control variables? *Research Report*, *16*(11), 866–870. doi:10.1111/j.1467-9280.2005.01628.x

Kuhn, D., & Franklin, S. (2007). The second decade: What develops (and how)? In W. Damon & R. M. Lerner (Eds.), *Child and adolescent development. An advanced course* (pp. 517–550). Hoboken, NJ: John Wiley & Sons.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *0*, 159–174. doi:10.2307/2529310

Lazonder, A. W., & Harmsen, R. (2016). Meta-analysis of inquiry-based learning: Effects of guidance. *Review of Educational Research*, *86*(3), 681–718. doi:10.3102/0034654315627366

Lehrer, R., & Schauble, L. (2000). Developing model-based reasoning in mathematics and science. *Journal of Applied Developmental Psychology*, *21*(1), 39–48. doi:10.1016/S0193-3973(99)00049-0

Lorch, R. F., Lorch, E. P., Freer, B., Calderhead, W. J., Dunlap, E., Reeder, E. C., … Chen, H.-T. (2017). Very long-term retention of the control of variables strategy following a brief intervention. *Contemporary Educational Psychology*, *51*, 391–403. doi:10.1016/j.cedpsych.2017.09.005

Lorch, R. F., Lorch, E. P., Freer, B. D., Dunlap, E. E., Hodell, E. C., & Calderhead, W. J. (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*(1), 18–35. doi:10.1037/a0034375

Mercer, N. (2013). The social brain, language, and goal-directed collective thinking: A social conception of cognition and tis implications for understanding how we think, teach, and learn. *Educational Psychologist*, *48*(3), 148–168. doi:10.1080/00461520.2013.804394

Mercer, N., Dawes, L., Wegerif, R., & Sans, C. (2004). Reasoning as a scientist: Ways of helping children to use language to learn science. *British Educational Research Journal*, *30*(3), 359–377. doi:10.1080/01411920410001689689

Nam, J., Choi, A., & Hand, B. (2011). Implementation of the science writing heuristic (SWH) approach in 8th grade science classrooms. *International Journal of Science and Mathematics Education*, *9*(5), 1111–1133. doi:10.1007/s10763-010-9250-3

Organisation for Economic Co-operation and Development. (2014). *Measuring innovation in education: A new perspective, educational research and innovation*. Author. doi:10.1787/9789264215696-en

Ozuru, Y., Briner, S., Kurby, C. A., & McNamara, D. S. (2013). Comparing comprehension measured by multiple-choice and open-ended questions. *Canadian Journal of Experimental Psychology*, *67*(3), 215–227. doi:10.1037/a0032918

Robertson, E. M. (2009). From creation to consolidation: A novel framework for memory processing. *PLoS Biology*, *7*(1), 11–19. doi:10.1371/journal.pbio.1000019

Schichow, M., Croker, S., Zimmerman, C., Höffler, T., & Härtig, H. (2016). Teaching the control-of-variables strategy: A meta-analysis. *Developmental Review*, *39*, 37–63. doi:10.1016/j.dr.2015.12.001

Schröder, E., Bödeker, K., & Edelstein, W. (2000). *The development of syllogistic reasoning. A manual including measurement properties and descriptive analyses*. Berlin: Max-Planck-Institut für Bildungsforschung.

Smit, J., Gijsel, M., Hotze, A., & Bakker, A. (2018). Scaffolding primary teachers in designing and enacting language-oriented science lessons: Is handing over to independence a fata morgana? *Learning, Culture and Social Interaction*, *18*, 72–85. doi:10.1016/j.lcsi.2018.03.006

Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics* (4th ed.). Needham Heights, MA: Allyn & Bacon.

Van Aalderen-Smeets, S. I., Walma van der Molen, J. H., & Asma, L. J. (2012). Primary teachers' attitudes toward science: A new theoretical framework. *Science Education*, *96*(1), 158–182. doi:10.1002/sce.20467

Van der Graaf, J., Segers, E., & Verhoeven, L. (2016). Scientific reasoning in kindergarten: Cognitive factors in experimentation and evidence evaluation. *Learning and Individual Differences*, *49*, 190–200. doi: 10.1016/j.lindif.2016.06.006

Van der Graaf, J., Segers, E., & Verhoeven, L. (2018). Individual differences in the development of scientific thinking in kindergarten. *Learning and Instruction*, *56*, 1–9. doi:10.1016/j.learninstruc.2018.03.005

Van de Sande, E., Verhoeven, L., Kleemans, T., & Segers, E. (in press). The linguistic nature of children's scientific reasoning. *Learning and Instruction*.

Verhallen, M., & Walst, R. (2001). *Taalontwikkeling op school. Handboek voor interactief taalonderwijs [Language development in school. Handbook for interactive language education]*. Bussum: Coutinho.

Vitale, M. R., & Romance, N. R. (2012). Using in-depth science instruction to accelerate student achievement in science and reading comprehension in grades -2. *International Journal of Science and Mathematics Education*, *10*(2), 457–472. doi:10.1007/s10763-011-9326-8

Wagensveld, B., Segers, E., Kleemans, T., & Verhoeven, L. (2015). Child predictors of learning to control variables via instruction or self-discovery. *Instructional Science*, *43*(3), 365–379. doi:10.1007/s11251-014-9334-5

Whitebread, D. (1999). Interactions between children's metacognitive abilities, working memory capacity, strategies and performance during problem-solving. *European Journal of Psychology of Education*, *14*(4), 489–507. doi:10.1007/BF03172975

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

Yeager, D. S., Henderson, M. D., Paunesku, D., Walton, G. M., D'Mello, S., Spitzer, B. J., & Duckworth, A. L. (2014). Boring but important: A self-transcendent purpose for learning fosters academic self-regulation. *Attitudes and Social Cognition*, *107*(4), 559–580. doi:10.1037/a0037637

Zimmerman, C. (2007). The development of scientific thinking skills in elementary and middle school. *Developmental Review*, *27*(2), 172–223. doi:10.1016/j.dr.2006.12.001