Res Sci Educ (2018) 48:301–324
DOI 10.1007/s11165-016-9569-z

The Effects of Video Feedback Coaching for Teachers on Scientific Knowledge of Primary Students

Sabine van Vondel[1] · Henderien Steenbeek[1] · Marijn van Dijk[1] · Paul van Geert[1]

Published online: 13 April 2017

Abstract The present study was aimed at investigating the effects of a video feedback coaching intervention for upper-grade primary school teachers on students' cognitive gains in scientific knowledge. This teaching intervention was designed with the use of inquiry-based learning principles for teachers, such as the empirical cycle and the posing of thought-provoking questions. The intervention was put into practice in 10 upper-grade classrooms. The trajectory comprised four lessons, complemented with two premeasures and two postmeasures. The control condition consisted of 11 upper-grade teachers and their students. The success of the intervention was tested using an established standardized achievement test and situated measures. In this way, by means of premeasure and postmeasure questionnaires and video data, an assessment could be made of the change in students' scientific knowledge before, during, and after the intervention. In this study, we primarily focused on the dynamics of students' real-time expressions of scientific knowledge in the classroom. Important indicators of the effect of the intervention were found. Through focusing on the number of explanations and predictions, a significant increase could be seen in the proportion of students' utterances displaying scientific understanding in the intervention condition. In addition, students in the intervention condition more often reasoned on higher complexity levels than

Electronic supplementary material The online version of this article (doi:10.1007/s11165-016-9569-z) contains supplementary material, which is available to authorized users.

✉ Sabine van Vondel
s.van.vondel@rug.nl

Henderien Steenbeek
h.w.steenbeek@rug.nl

Marijn van Dijk
m.w.g.van.dijk@rug.nl

Paul van Geert
p.l.c.van.geert@rug.nl

[1] Department of Developmental Psychology, University of Groningen, Grote Kruisstraat 2/1, 9712 TS Groningen, the Netherlands

students in the control condition. No effect was found for students' scientific knowledge as measured with a standardized achievement test. Implications for future studies are stressed, as well as the importance of enriching the evaluation of intervention studies by focusing on dynamics in the classroom.

Keywords Scientific knowledge · Teaching intervention · Upper-grade primary school students · Video feedback coaching

Introduction

Scientific knowledge is considered increasingly important in order for future citizens to be able to fully participate in society (Esmeijer and Van der Plas 2012; Silva 2009). This consideration translates into the need for students to learn scientific subject matter and acquire understanding of scientific concepts in the earlier stages of their school careers (Bybee and Fuchs 2006; Jorde and Dillon 2013). However, a concerning decline has been observed when it comes to students' interest in studying science and technology-related disciplines during primary education (Blickenstaff 2005; Dekker et al. 2008; Fouad et al. 2010). Recently, inquiry-based learning has been a topic of interest in many studies concerning effective science education (e.g., Furtak et al. 2012). One goal of science-educational interventions in primary education is to influence the teaching strategies of teachers in such a way that these lead to cognitive gains of students in scientific knowledge. Several teaching interventions involving inquiry-based learning principles have been designed to support teachers in the implementation of successful science education (e.g., Borman et al. 2008; Cuevas et al. 2005; Penuel et al. 2011; Şimşek and Kabapinar 2010; Wetzels 2015).

The aim of this study is to investigate the effects of a professionalization intervention on students' scientific knowledge, as expressed by the quantity and complexity of their utterances during science lessons. The intervention was a video feedback coaching program for teachers (VFCt) that was conducted in the upper grades of primary education and was aimed at improving teacher-student interactions to stimulate students' scientific learning during inquiry-based activities. In this study, we primarily focused on the dynamics of students' real-time expressions of scientific knowledge in the classroom.

Learning Science

Inquiry-based teaching is a pedagogical method that is frequently used in science education (Minner et al. 2010). It is a way of facilitating students' active participation in learning science. In other words, it is a way to engage them in scientific processes and invite them to use critical thinking skills as they search for answers (Gibson and Chase 2002). Note that inquiry-based learning and inquiry-based teaching are related, in that inquiry-based teaching provides opportunities for students to learn in an inquiry-based manner. Inquiry-based learning is a student-centered, active learning approach comprising hands-on activities and higher-order thinking skills (Cuevas et al. 2005; Oliveira 2009; Tang et al. 2009). The role of the teacher is to facilitate the learning process by encouraging students to verbalize ideas and thoughts and by asking higher-order thinking questions (Lee and Kinzie 2011; Oliveira 2009; Van Zee et al. 1997). The empirical cycle (De Groot 1994) is often used as a means to structure the acting and thinking process of the students (White and Frederiksen 1998). The empirical cycle, consisting of hypothesizing, designing experiments, observing, and explaining, should be understood as a method that helps teachers to accompany students in their processes of exploration and inquiry.

Hands-on activities have been shown to trigger enthusiasm and surprise (Bilgin 2006). Although students' enthusiasm is an excellent starting point, learning can only take place when students are challenged to think and communicate ("minds-on") about the hands-on activities (Van Keulen and Sol 2012). The experience alone of doing a scientific experiment does not coincide with learning gains (Lutz et al. 2006). An important goal of inquiry-based teaching is thus preparing students to engage in scientific reasoning, such as solving problems, formulating predictions, developing explanations, drawing conclusions, being creative, making decisions, and extending knowledge to new situations (Klahr 2000; Cito 2010; Kuhn 2010; Mullis et al. 2011). The combination of hands-on activities and stimulation of scientific reasoning skills is what makes inquiry learning an effective way of teaching science. All of these activities will contribute to the development of students' scientific knowledge.

Different Forms of Scientific Knowledge

Science is a term that is used to describe both a body of content knowledge and the activities that give rise to that knowledge, such as scientific reasoning (Zimmerman 2000). Miller and Hudson (2007) distinguish three types of knowledge: procedural, declarative, and conceptual knowledge. All three types are important for cognitive development, as declarative knowledge provides the foundation for procedural knowledge and procedural knowledge interactively develops with conceptual knowledge (Miller and Hudson 2007). In this study, the focus is on declarative and conceptual knowledge.

Declarative knowledge refers to recall of factual, often domain-specific, information (Scardamalia and Bereiter 2006). Gains in declarative knowledge consist of adding new information to existing schemas (Miller and Hudson 2007). This form of knowledge increases as students become older and is the product of a significant amount of practice and attention.

Conceptual knowledge refers to an *understanding* related to underlying principles, which, in the context of science education, is also known as scientific understanding (e.g., Meindertsma 2014). It requires general scientific abilities, one of which is scientific reasoning (Bao et al. 2009; Zimmerman 2000), to connect bits of knowledge into an integrated whole (i.e., to assimilate new information into existing schemas or to accommodate the existing schemas to new information). Scientific reasoning skills foster the acquisition of deeper and more complex understanding of scientific content matter. The development and application of such capacities in the classroom will prepare students to design their own investigations to solve scientific, engineering, and social problems in the real world (Bao et al. 2009; NGSS Lead States 2013). Students' scientific understanding comes to the forefront in their actual activities while performing an experiment (non-verbal) and in their verbal utterances (such as predictions and explanations, which are the focus of this study) during science activities. These skills can be captured by focusing on how they emerge in interaction with the context (Guevara et al. in preparation; Rappolt-Schlichtmann et al. 2007; Van der Steen et al. 2012).

Assessment of Scientific Knowledge

It is important to determine whether educational interventions offer cognitive learning gains for students, i.e., whether they offer a combination of new and/or improved declarative, conceptual, and procedural knowledge. Several studies have shown that students' display of predictions and explanations are associated with cognitive gains (Christie et al. 2009; Thurston et al. 2008; Webb 1989). Research into the cognitive gains of students in science may be classified into two main streams: studies that use static measurements (i.e., one measurement from each student in the group with the results based on aggregated scores) and studies oriented toward more intensive observational studies of students learning in situ (dynamic or process approach). Both approaches have improved our understanding of students' learning gains in science, but how the two assessments relate to each other is relatively unknown.

The static measurement can be used to analyze the global effect of the intervention, for instance, by focusing on the difference in score on a science test between premeasure and postmeasure, in which half of the participants receive an intervention while the other half do not (control condition). These are usually relatively large-scale studies in which students are grouped according to general characteristics, such as gender, level, and condition. The cognitive gains are typically assessed by means of standardized science and mathematics tests, at the end of the academic year in case of academic progress and at premeasure and postmeasure in case of an intervention (e.g., Borman et al. 2008; Penuel et al. 2011; Şimşek and Kabapinar 2010). Assessment in science education often emphasizes factual recall—declarative knowledge—by focusing on specific outcomes of the learning process (Bao et al. 2009). These studies give an idea of global developmental trends across cohorts and/or conditions. For instance, Trends of International assessment Mathematics and Science Study (TIMSS; Mullis et al. 2011) is an international instrument that focuses on both declarative knowledge and scientific understanding of mathematics and science.

The more dynamic process approach, on the other hand, can be used to assess the process of learning over time (e.g., multiple activities or lessons). This second form of assessment is often based on more intensive experimental or observational studies of students' learning in the actual context of learning (e.g., Meindertsma 2014; Van der Steen et al. 2012). One example of this is an assessment based on the actual production of explanations and predictions in the context of a real science activity (Meindertsma 2014; Rappolt-Schlichtmann et al. 2007). The goal of this assessment is to examine the capabilities of students as they interact with the teacher *during* science activities. This means assessing how scientific knowledge emerges in interaction with the teacher (as the knowledge is expressed in interaction, this will be termed *situated* knowledge). Situated declarative knowledge can for instance be assessed by the teacher—in interaction—by asking knowledge-based questions (Oliveira 2009). Situated scientific understanding can be assessed in the classroom by investigating students' scientific reasoning skills (Rappolt-Schlichtmann et al. 2007). The complexity level of situated scientific understanding of students can be mapped by focusing on their predictions and explanations during scientific experiences (e.g., Meindertsma 2014; Rappolt-Schlichtmann et al. 2007; Van der Steen et al. 2014). Predictions and explanations are considered to be distinct emergent processes, which means that predictions and explanations emerge in interaction with the context and are not necessarily intrinsically related or originating from one single mental representation (Meindertsma 2014).

A framework for determining complexity levels of students' scientific knowledge is given by dynamic skill theory. Dynamic skill theory (Fischer 1980) is a cognitive developmental theory focusing on how skills are constructed in specific domains and in specific (transient) contexts. Hence, science learning is seen as a process of co-construction that emerges in real time and develops over multiple interactions (Thelen and Smith 1994). This means that students' scientific knowledge emerges out of the interaction between teacher and students. The level of the skills captures the complexity of children's reasoning from moment to moment

(Meindertsma 2014; Rappolt-Schlichtmann et al. 2007; Van der Steen et al. 2014). The development of students' scientific understanding can be categorized into three different tiers. These tiers should not be visualized as a static ladder but like a web (Fischer and Bidell 2006) in which development of students is becoming increasingly differentiated. Each tier consists of three different levels in which each consecutive level is more complex and includes each underlying level within that tier. Each tier, however, is qualitatively different from the other tiers (Fischer 1980).

The first tier consists of sensorimotor observations and explanations, which implies that simple observable connections are given. A student might explain that "the oil floats when *you* put it in there," for example. The second tier comprises representational predictions and explanations, which means that students use higher-order thinking skills to go beyond simple perception-action couplings. Students understand that an object has a specific characteristic, independent of the present situation. They can, for instance, make a prediction about what is going to happen when the salt is put into the water/oil fluid without directly seeing it. The third tier constitutes abstract explanations: students are capable of generalizing ideas about the object outside specific situations. A student might for instance explain that "molecules in the water are strongly drawn towards each other… probably leading to surface tension… the water and oil cannot blend because of that" or "the density of the water is higher compared to the density of the oil, and fluid with lower density floats."

To conclude, both static and process measurements aim to capture cognitive gains of students during and/or after science lessons. A difference is that static measurement focuses on their product as captured in a single score for a large sample of students, where students need to answer questions outside the regular learning situation. The focus is on what a student has learned. In contrast, dynamic measurement can be used to assess students' progress by focusing on the moment-to-moment behaviors (e.g., their verbal utterances as elicited by the teacher within the regular learning situation during a science class). The focus is on what a student learns in authentic situations, with support of the context. This provides insight into what the student's optimal level of performance is (Fischer 1980), which is typically the level that predicts learning (Fischer and Van Geert 2014). The present study will assess the effectiveness of a VFCt intervention using both measures in order to gain a more comprehensive evaluation of the learning gains of the students.

Video Feedback Coaching

Video feedback is a powerful form of training that can enhance or change teachers' skills (see also Hintze and Matthews 2004; Mortenson and Witt 1998; Seidel et al. 2011; Wetzels 2015). An effective coaching program combines classroom observation, micro-teaching, video feedback, and practice in the classroom (Wade 1984) with goal setting at the beginning of a coaching intervention. A promising method for implementing evidence-based instructional strategies (i.e., establishing behavioral change in teacher behavior) is that of providing feedback on real-time behavior (Noell et al. 2005; Reinke et al. 2009). Immediate video feedback is very beneficial for learning (Fukkink et al. 2011), and the effect of feedback is best when a 3:1 positivity ratio is used (Fredrickson 2015). This means that reflection is based on three fragments that show positive (examples of) teacher behavior and one fragment that shows (an example of) teacher behavior that might be improved in the future. Several scholars (Fukkink et al. 2011; Hargreaves et al. 2003; Kennedy et al. 2011) have demonstrated that employing video feedback methods has positive effects on teachers' use of stimulating behavior, sensitive responsivity, and verbal stimulation.

The aim of the intervention is to improve the teacher's skills in supporting students to develop more complex cognitive skills. As the aforementioned evidence shows, the best way of improving teachers' skills is by entering the context of their actual teaching practice and making use of their real-time activities as captured by means of video recordings.

Present Study

The aim of this study is to investigate the effect of the VFCt professionalization intervention on students' scientific knowledge. The intervention contained the following key goals: (1) improving teachers' knowledge and skills in order to be better able to teach science, (2) encouraging behavioral change by means of video feedback coaching. Teachers were encouraged to use thought-provoking questions to spark students' display of situated scientific knowledge and to use the steps of the empirical cycle (De Groot 1994; White and Frederiksen 1998) to guide the acting and thinking process of students. These instructional strategies were aimed at increasing students' participation and, in doing so, helped them develop more and higher complexity levels of scientific understanding.

The intervention will be approached from different angles in order to provide insight into its effect on students' cognitive gains as measured by standardized learning achievement tests and more dynamic measures of cognitive gains.

The following questions will be examined:

1. To what extent does the students' knowledge—as measured with the standardized science assessment—change following the intervention, and how does this compare to changes in the control condition?
2. To what extent does the students' *situated* scientific knowledge—as measured by their scientific reasoning skills during actual science lessons—change following the intervention and how does this compare to changes in the control condition?

The assumption is that an observable increase in students' scientific knowledge can be brought about by improving the skills of the teacher in stimulating the skills that are needed by the students in order to develop scientific knowledge. The Conclusion and Discussion section will deal with the question of how the more static and more dynamic measurements relate to one another.

Method

Participants

Twenty-three upper-grade teachers (Grades 3–6, which in the Dutch school system is Groups 5–8) and their students (586 in total) participated during the school years 2013-2014 and 2014-2015 (see Table 1). All participating teachers taught at primary schools in the north of the Netherlands. The average age of the teachers in the intervention condition was 38 (range 23–54), with an average teaching experience of 13 years (range 1–32). The students had an average age of 10.7 (range 8.4–13.2; 51% boys). The teachers in the control condition were comparable to the teachers in the intervention condition on the basis of age (M = 37 years; range 24–54), teaching experience (M = 13 years; range 1–30), and which grades they taught (see Table 1). In the control condition, the students' average age was 10.6 (range 7.3–13.2; 53% boys). The teachers and parents of each participating student gave active consent before the start of the study. The study was approved by the Ethical Committee Psychology of the University of Groningen (the Netherlands).

Procedure

The teachers were recruited using flyers and personalized e-mails, which were first sent to school boards, followed by a telephone call to ask whether teachers could be contacted in person. All teachers were interested in teaching science as they voluntarily enrolled into the program to improve their inquiry-based teaching skills.

A quasi-experimental pre-test-post-test control condition design was used to assess the effects of the intervention. The teachers who participated in the intervention condition were offered the VFCt, while the control condition functioned as a waiting list condition. This means that the teachers who participated in the control condition were offered the opportunity to participate in a free VFCt after the control intervention. Alternatively, they could choose to receive a workshop for their team or a gift card for classroom materials.

Table 1 Participants' details

Teacher	Condition	Grade	Gender	Age	Teaching experience	Number of students
1	Intervention	3/4	Male	49	7	21
2	Intervention	4	Female	30	10	25
3	Intervention	4	Male	54	32	31
4	Intervention	5	Female	23	1	26
5	Intervention	5	Female	34	12	28
6	Intervention	5/6	Female	50	15	15
7	Intervention	5/6	Female	29	7	24
8	Intervention	5/6	Female	39	17	19
9	Intervention	6	Male	46	15	22
10	Intervention	6	Female	27	5	35
11	Control	3/4	Female	42	20	25
12	Control	3/4	Female	40	18	22
13	Control	4/5	Male	29	4	27
14	Control	4/5/6	Male	33	8	29
15	Control	5	Male	35	12	19
16	Control	5	Female	35	8	14
17	Control	5	Female	30	1	14
18	Control	5/6	Female	46	24	27
19	Control	5/6	Female	26	4	29
20	Control	6	Female	54	30	24
21	Control	6	Male	NA	NA	16

Two teachers were excluded for analysis (see Implementation Criteria section)

Video Feedback Coaching Program for Teachers[1]

Each teacher in the intervention condition was observed during eight science and technology lessons within a period of approximately 3 to 4 months (Fig. 1). The first two lessons were pre-intervention lessons. In between the second and third lesson, the teacher received an introduction about the aims of the professionalization intervention. During this educational introduction, information about inquiry-based instruction strategies was provided and discussed. This included information about the use of the empirical cycle (De Groot 1994), thought-provoking questioning styles (Chin 2006; Oliveira 2010), scaffolding (Van de Pol et al. 2010), and inquiry-based learning activities (Gibson and Chase 2002). Several video fragments of teacher-student interactions were shown to illustrate the importance and effect of high-quality interactions during science and technology activities. During the intervention-stage (Lessons 3 to 6), video feedback coaching was given weekly after every science and technology lesson. The video feedback coaching took place immediately after each lesson. During the lesson, the coach (first author) selected several fragments to critically reflect upon, of which three were positive fragments and one was a fragment that showed an example of teacher behavior that could be improved. The teacher and the coach discussed the video recordings of the lessons, making use of the teacher's personal learning goal and focusing on the teacher's interaction with the students. Attention was paid to inquiry-based instruction strategies, such as eliciting remarks, and teachers were encouraged to show more of these skills in their next lessons. The final two lessons were post intervention and were videotaped approximately 2 months (range 4–14 weeks) after the last video feedback coaching session. To assure teaching-as-usual conditions, teachers were free to choose whether or not they continued providing science and technology lessons between the intervention and post-measures.

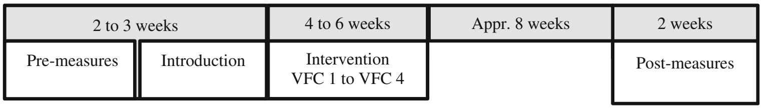
The control condition teachers were observed during four science and technology lessons that were taught as usual, in other words, without receiving the intervention. The teachers taught two lessons at the beginning of the intervention, (pre-measures), and two lessons approximately 2.5 months later (range 7–17 weeks; post-measures). The teachers were not given any teaching instructions in the period between the measures. They were free to choose whether or not they would continue to provide science and technology lessons between the premeasures and postmeasures. At the postmeasure stage, the teachers completed a questionnaire that included a question about how often they teach science and technology. Approximately 65% said that they teach science and technology on a regular basis (between once a week and at least three times per month).

Implementation Criteria

According to Durlak and DuPre (2008), the effect of an intervention program depends, among other things, on the quality of the program implementation (how well the program components have been carried out), the quantity of the program implementation (how much of the original program has been delivered), and participant responsiveness (the degree to which the program stimulates the interest or holds the attention of participants). As the following section will demonstrate, our intervention program successfully met these three requirements.

To ensure the *quality* of the program, a trained coach was responsible for the implementation. That is, the same coach provided all participants with identical information during the

[1] For a more elaborate description of the intervention, the reader is referred to Appendix.



2 to 3 weeks		4 to 6 weeks	Appr. 8 weeks	2 weeks
Pre-measures	Introduction	Intervention VFC 1 to VFC 4		Post-measures

Fig. 1 Design of the study for the intervention condition. Note that for the control condition, there was a comparable amount of time between premeasures and postmeasures

introductory session, videotaped all lessons—using a camcorder with extended microphone—and was responsible for the guided reflection after each lesson. In addition, the coach made use of prespecified guidelines for each coaching session.

Another important element is the timeframe in which the intervention was conducted. In order to be able to compare classrooms, it was considered important to implement all the coaching during the same part of the school year. In addition, it was important that the period of implementation comprised a similar number of weeks for each classroom.

Although the intervention was intended as adaptive support, some standardization was implemented during data collection. The teachers were asked to provide all lessons using a commonly used teaching format: the direct instruction model. According to this model, the teacher starts with a plenary introduction by taking an inventory of existing knowledge and introducing the topic of the day, followed by a middle part in which students work on their own or in groups and the teacher facilitates the learning process by asking questions. The lesson ends by returning to a plenary conclusion or discussion. In addition to these requirements, the teachers were asked to spend each lesson on the “earth and space” system, teaching topics such as weather, air pressure, gravity, and the positions of the moon.

As the number (*quantity*) of components of the originally intended program is essential for the implementation, teachers were only included in the intervention condition if they met the essential elements of the coaching, i.e., participation in the educational session and in all video feedback coaching sessions (within a period of 6 weeks). This meant exclusion of one teacher who was not able to attend the coaching session of the first lesson and who did not provide a fourth lesson. Teachers from the control condition were included if they met at least 75% of the requirements (i.e., if they participated in at least three out of four measurements). Again, one teacher was excluded because he only provided the pre-measure lessons. This resulted in a total of 10 teachers who participated in the intervention, while the other 11 teachers were part of the control condition.

The *participant responsiveness* was considered substantial, as the teachers in the intervention condition all actively participated during the educational session and the coaching sessions. Their active participation was further observable in their enthusiasm to design new lessons, their willingness to make time for the coaching sessions, and their eagerness to work on and reformulate their learning goals—all of which was done with the intention to provide science lessons that are stimulating for their students.

Measures and Variables

Students from both the intervention condition and control condition completed the same paper-and-pencil test just before the pre-measure and just before the post-measure. Students’ expressions of scientific knowledge when interacting with the teacher were captured by means of video measures during the eight science lessons. The chosen unit of coding was the utterance. Utterance boundaries were determined on the basis of turn-taking and pauses.

Four segments from the central section of each lesson were selected for coding: a 3-minute segment from the beginning, a 3-minute segment from the end, and two 2-minute segments from the middle. The central section was defined as the part of the lesson where students were working and the teacher walked around to see whether they needed assistance. To ensure that the starting point of the central section would be the same for all lessons, the moment of the first substantive verbal expression was used. The two 2-minute segments were moments in which a great deal of teacher-student interaction took place, and the final 3-minute segment took place before the teacher resumed lecturing at the front of the classroom. This procedure led to the generation of a time series of 10 minute per lesson consisting of teacher and student utterances and portions of time during which no utterances occurred.

The students' *scientific knowledge* was assessed using an established and standardized subtest of the Final Test for Primary Education: World orientation, part 1: Nature education (Cito 2010). This test is usually administered in Grade 6 at the end of the academic year. The test consists of 30 multiple-choice questions, each question having four possible answers. Ninety percent of the questions covered scientific domains such as biology, chemistry, physics, and technology, while 10% focused on questions concerning scientific understanding. Scores were based on the number of correct answers.

The students' *situated scientific knowledge* was assessed on the basis of the video recordings. Coding was done in three phases. During Phase 1, student utterances were classified in the following categories: "task-related complex utterances," "task-related non-complex utterances," or "non-task-related utterances." Non-task-related utterances are utterances that are clearly unrelated to the task at hand, for instance: "Look, my mother is walking outside." Task-related non-complex utterances consisted of procedural utterances: utterances about observations and answers to closed questions related to the task at hand. These are utterances that are task-related but could not be scored on the complexity of understanding. In this study, we focused on task-related complex utterances, representing the complexity of students' utterances. This category consisted of expressions of *situated knowledge*, *predictions*, and *explanations* (see Table 2 for examples). Note that both correct and incorrect explanations were coded. Incorrect explanations can be coded at the same level as correct ones since the complexity level is equal. In this study, we reasoned, following Schwartz and Fischer (2004), that misconceptions can be seen as a different pathway that can lead to a higher level of scientific knowledge.

Phase 2 consisted of quantifying the task-related complex utterances. Students' utterances were quantified based on their complexity, using a scale that is based on skill theory. Students' utterances were first scored for complexity, using a 10 point scale (Table 2). Here, a *1* means the least complex utterance on a sensory motor level (e.g., an expression of what a student sees for instance, "It [the balloon] is white"), while *10* could be scored if a student expressed understanding about global laws and principles (e.g., the abstract principles of thermodynamics can be applied to the situation at hand).

Phase 3 consisted of extracting predictions and explanations from the task-related complex utterances, as predictions and explanations can be used as measures for scientific understanding. Since predictions and explanations are considered to be distinct emergent processes, they are analyzed separately. The lowest possible level for explanations was Level 3 (sensorimotor system), since an explanation contains a combination of a description of what happened and a reason why it happened. The lowest possible level for predictions was Level 4, since a prediction requires a representation of the situation. Ten-to 12-year-olds are expected to be capable of reaching the seventh level of understanding (Fischer and Bidell 2006), where they

Table 2 Description and examples of the complexity levels as expressed in students' verbal utterances

Complexity level	Description	Example
Sensorimotor	A student expresses:	
1 Sensorimotor actions	A single characteristic of the task	It [the paper] is *white*.
2 Sensorimotor mappings	A link between single task characteristics or simple comparisons	It is *white* and that one is *yellow*.
3 Sensorimotor system	An observable causal relation	The paper collapses *because I blow*.
Representational		
4 Single representations	One part of the explaining mechanism; not directly observable relations or objects and simple predictions	*I think* the paper will ascend.
5 Representational mappings	Two or more parts of the explaining mechanism. Predictions in terms of a relation between two single representations. A superlative in one representation (but not yet linked to the change in another representation)	Because I blow, there is *more space* and then the *paper can go down*.
6 Representational systems	A combination of all relevant parts of the explaining mechanism; a coupling between two representational mappings, i.e., a change in one representation causes a change in another	Because I blow hard*er*, the paper can drop low*er*.
Abstraction		
7 Single abstractions	A general (immaterial) concept that goes beyond (representations of) the material	Because the balance is gone… the air beneath the paper is pushed away [when I blow]… so… the *air pressure* drops down, but the pressure above the paper remains the same and thus pushes the paper down

In this study, with young students, the highest complexity levels will not be reached. Therefore, only levels 1 to 7 are presented

are able to express abstract thinking skills (e.g., relate abstract concepts to the situation at hand, such as in the utterance: "the air pressure pushes the paper towards the table").

Coding was done by means of the program Mediacoder 2009 (Bos and Steenbeek 2009). In order to establish inter-observer reliability for the application of the coding scheme, the inter-observer agreement was determined in advance by the first author and several independent coders for each phase. Table 3 shows that the inter-observer agreement was considered substantial.

Table 3 Percentage agreement with corresponding Cohen's kappa for each phase of the coding

	% Agreement (range)	Kappa
Complex or not	87 (76–97%)	0.82
Tier	76 (64–82%)	0.78
Skill level	74 (58–79%)	0.77

Data Analysis

Students' Scientific Knowledge: Pre-measures and Post-measures

Data obtained via the student tests were analyzed using SPSS 23.00. In order to investigate the effects of the intervention on students' scientific knowledge, a 2 (condition: control vs. intervention) × 3 (grade: 4 vs. 5 vs. 6) × 2 (time: premeasure vs. postmeasure) × 1 (scientific knowledge) repeated measures ANOVA was conducted, with condition and grade as between-subjects factors, time as a within-subjects factor, and scientific knowledge as dependent variable.

Students' Situated Scientific Knowledge: Pre-measures, Intervention Measures, Post-measures

For further data analysis, the classroom of students as a whole was taken as the unit of analysis. As the collected data consisted of a small group of classrooms, a permutation test (Todman and Dugard 2001) was used. Due to the size of the group, the dependency between measures, and the fact that there were multiple measures, the assumptions of traditional statistics (such as a paired sample *t* test) could not be met. Therefore, a non-parametric test (Monte Carlo analyses) was used to test whether differences in the complexity of students' level of scientific knowledge over several lessons were equal to or smaller than what could be expected on the basis of change. The random permutation test using Poptools was used to test the empirical results in relation to a statistically simulated baseline of random patterns (Hood 2004). This means that the non-parametric test statistically simulates the null hypothesis that the observed probability of the relationship or property is based on chance alone. For instance, to examine whether the complexity level at pre-measure significantly differs from the complexity level at postmeasure for the intervention condition, the difference score of the empirical data is calculated. Next, the complexity level of all intervention classrooms were randomly shuffled over premeasure and postmeasure (values were randomly drawn from the data without replacement), and each time the difference score was calculated. This random shuffling, i.e., data generated on the basis of the null hypothesis model that there was no effect of the intervention, was permutated 10,000 times in order to calculate whether the empirically found difference between pre-measure and post-measure could be expected to occur on the basis of chance. Note that in a similar vein complexity levels of each condition can be used to examine differences between conditions.

The random permutation test provides an estimation of the exact *p* value, and the greater the number of permutations, the closer this estimation comes to the exact value (see Gigerenzer 2004; Schneider 2015). As significance scores are not directly linked to practical significance (Sullivan and Feinn 2012), the effect size was calculated using Cohen's *D*. Following Sullivan and Feinn, an effect size of 0.2 is considered small, 0.5 medium, 0.8 large, and 1.3 or higher very large. Results with a *p* value smaller than 0.05 and an effect size greater than 0.8 are understood as convincing evidence strongly supporting our belief that the intervention was effective. Results with a *p* value between 0.05 and 0.1 and an effect size that is medium to large can be understood as less convincing evidence, providing weak support to our belief that the intervention was effective. Results with small to very small effect sizes—whatever the *p* value—can be understood as unconvincing evidence, providing no support to our belief that the intervention was effective. In addition, the 95% confidence interval, which is based on the standard deviations, was displayed for each analysis. Note that the data of the pre-measures are taken together as one measure in the analysis, and the same applies for the post-measures. This was done to counteract random hits. Furthermore, for the intervention condition, the graphs display the intervention lessons in order to highlight how the change from premeasure to postmeasure occurs.

The analysis was done in three steps, and it is by focusing on these three dimensions that the research question will be answered. The focus was on the proportion and number of complex utterances (quantity) and the average level of complex utterances, as they can all be considered indicators of learning gains. Each step in the analysis is accompanied by a descriptive analysis that gives information about individual pathways, such as the percentage of individuals who showed at least the average finding. This refers those individual pathways that at least doubled the number of complex utterances between pre-measure and post-measure. Firstly, the proportion of complex utterances relative to the "total number of utterances" (complex utterances + task-related utterance but non-complex utterances + non-task-related utterances) was calculated for each classroom. The average proportion in the intervention and control conditions was used to determine whether and to what extent the frequency of complex utterances increased as a result of the intervention. Secondly, the number of predictions and explanations were calculated for each classroom in order to study what the details of the change in the proportion of complex utterances were. In order to further refine the findings, the analysis zoomed in on the higher complexity levels (4, 5, 6, and 7) to determine what cognitive levels mainly accounted for the change in the students' scientific understanding. Thirdly, the average complexity level of the explanations and predictions was calculated in order to find out whether the complexity level of the students' scientific understanding had increased over time. In each step, a Monte Carlo permutation test as described above was used to test both the difference between the premeasures and postmeasures and the difference between the conditions.

Results

Students' Scientific Knowledge on the Achievement Test

A difference was found between the grades over time on the achievement test ($F(2, 277) = 4.00$, $p = 0.02$, partial $\eta^2 = 0.03$), but no effect was found for condition over time ($F(1, 277) = 0.27$, $p = 0.60$, partial $\eta^2 = 0.001$). However, the result of the difference between the grades shows a small effect size. Therefore, the difference is classified as unconvincing evidence, providing no support to our belief that the intervention was effective for students' cognitive gains.

Students' Situated Scientific Knowledge

Proportion of Complex Utterances

Descriptives The task-related complex utterances category, as measured in all groups and during all lessons, was composed of approximately 50% explanations, 35% predictions, and approximately 15% *declarative knowledge* statements. For both conditions, the complex utterances were mostly in the representational tier (M = 86% of the utterances; range 79–

89%), followed by the sensorimotor tier (M = 10 %; range 7–17%), while the least number of utterances were in the abstract tier (M = 4 %; range 3–5 %). Preliminary analysis showed that the utterances in the declarative knowledge category were, on average, displayed once per lesson per classroom. This category is therefore excluded from further analysis.

Analysis Figure 2 gives no indication that there was a difference between the intervention and control condition when it comes to the average pre-measure of the proportion of task-related complex utterances (intervention = 18%, control = 16%, $p = 0.29$, $d = 0.26$, small effect size). In addition, the probability that the difference between the pre-measure and post-measure of the proportion of complex utterances was based on chance alone is very low for the intervention condition, but considerable for the control condition (intervention: M_{pre} = 18% vs. M_{post} = 28%, $p = 0.02$, $d = 1$; control: M_{pre} = 16% vs. M_{post} = 21%, $p = 0.15$, $d = 0.5$). When looking at the average postmeasure, the intervention condition showed more complex utterances than the control condition (intervention = 28%, control = 21%, $p = 0.12$, $d = 0.5$). This means that the intervention condition and control condition started the intervention with a roughly similar number of complex utterances. A difference over time was found for the intervention condition, following a gradual pathway of change, as well as a difference between the conditions at postmeasure. For the intervention condition, the graph displays how the change from pre-measure to post-measure occurs.

Number of Complex Utterances

Descriptives At postmeasure, classrooms in the intervention condition had almost doubled their number of task-related complex utterances compared to premeasure (Nr_{pre} = 266 vs. Nr_{post} = 509). Figure 3 shows that this doubling took place after the intervention. This group finding was visible in 60% of the individual classrooms (for whom the number of complex utterances at least doubled between premeasure and postmeasure), while 20% showed no change and 20% showed a decrease. The control condition did not show such an increase in complex utterances (Nr_{pre} = 264 vs. Nr_{post} = 270). This group finding was visible in 63% of the

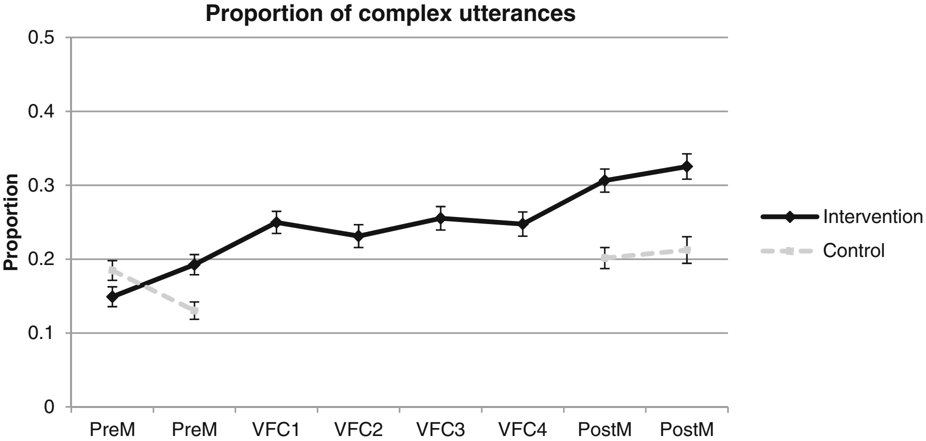


Fig. 2 Proportion of complex utterances and 95% confidence intervals, relative to the total number of student utterances, for intervention condition and control condition

individual classrooms (the number of complex utterances remained constant); while 27% showed a positive change and 9% seemed to show a decrease.

Analysis Figure 3 depicts the average number of *explanations* per condition per lesson. At pre-measure, students of both the intervention and the control conditions uttered approximately seven explanations ($I_{ntervention} = 6.4$ vs. $C_{ontrol} = 6.8$, $p = 0.47$, $d = 0.05$). At post-measure, however, the classrooms in the intervention condition doubled the number of explanations compared to the pre-measure ($M_{pre} = 6.4$ vs. $M_{post} = 13.5$, $p = 0.05$), while the control classrooms remained rather constant ($M_{pre} = 6.8$ vs. $M_{post} = 5.5$, $p = 0.65$). As shown by the Monte Carlo analysis, there is only a very small chance that this difference between the conditions at the post-measure was purely based on chance ($I_{ntervention} = 13.5$ vs. $C_{ontrol} = 5.5$, $p = 0.02$, $d = 0.85$).

Figure 4 depicts the average number of predictions per condition. At pre-measure, the intervention and the control conditions uttered roughly the same number of predictions ($I_{ntervention} = 4.1$ vs. $C_{ontrol} = 3$, $p = 0.2$, $d = 0.39$). The intervention condition shows a peak at the start of the intervention. A difference was found at post-measure, however, namely that the classrooms in the intervention condition uttered more predictions compared to the control classrooms ($I_{ntervention} = 8.1$ vs. $C_{ontrol} = 4$, $p = 0.06$, $d = 0.7$).

The representational tier was found to be the most common complexity level and will be further explored. At pre-measure, both conditions uttered approximately the same number of utterances at the representational level ($I_{ntervention} = 209$ vs. $C_{ontrol} = 224$, $p = 0.21$). The intervention condition showed an increase in the number of representational utterances ($Nr_{pre} = 209$ vs. $Nr_{post} = 457$, $p < 0.01$), while for the control condition, the number of representational utterances remained rather constant ($Nr_{pre} = 224$ vs. $Nr_{post} = 242$, $p = 1$). As shown by the Monte Carlo analysis, there is only a very low probability that this difference in the intervention condition was purely based on chance, while this probability is very high for the control condition.

A more fine-grained analysis showed that the number of *single representations* (Level 4) tremendously increased for the intervention condition between pre-measure and post-measure ($I_{pre} = 170$ vs. $I_{post} = 388$, $p < 0.01$; $C_{pre} = 180$ vs. $C_{post} = 206$; $p > 0.99$). In addition, the higher complexity levels of scientific reasoning (5, 6, 7) were uttered approximately the same number

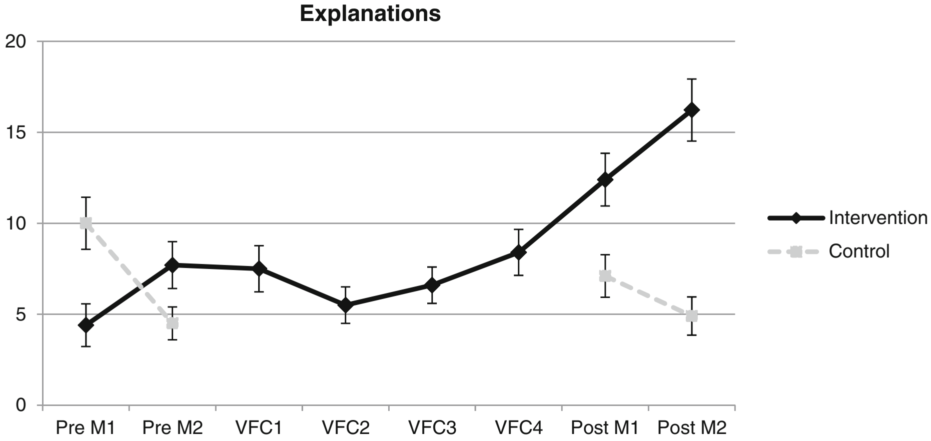


Fig. 3 Average number of complex explanations per lesson for intervention condition and control condition

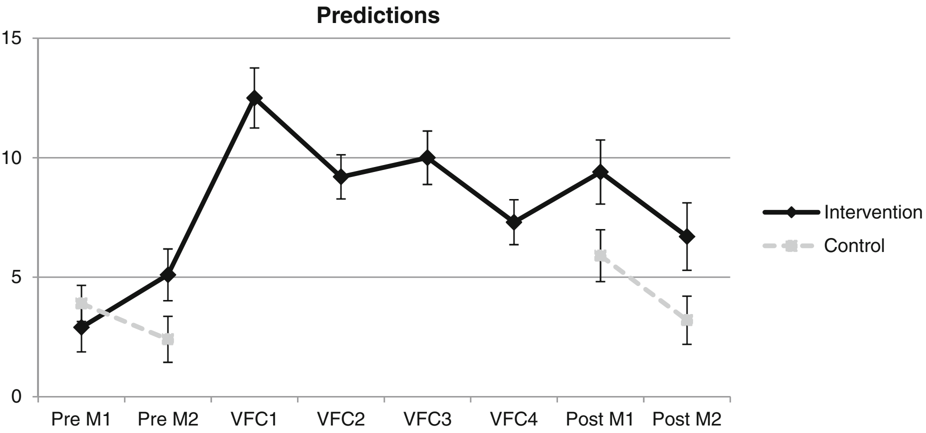


Fig. 4 Average number of complex predictions per lesson for intervention condition and control condition

of times in both conditions at pre-measure ($I_{ntervention} = 55$ vs. $C_{ontrol} = 69$; $p = 0.12$), while the intervention condition manifested more higher complexity levels at post-measure ($I_{pre} = 55$ vs. $I_{post} = 103$, $p < 0.01$; $I_{ntervention} = 103$ vs. $C_{ontrol} = 44$; $p = 0.04$).

Average Complexity Levels

Descriptives The complexity levels of predictions ranged from 4 to 5 in both conditions. The average complexity level of predictions in the classrooms in both conditions seemed to be constant in all classrooms over all lessons (M = 4.08; range 4 to 4.3). The complexity levels of explanations ranged from 3 to 7 in both conditions. The average complexity level of explanations in the classrooms ranged from 3.3 to 5 over all lessons in the intervention condition and from 3 to 5.1 over all lessons in the control condition.

Analysis The difference found between the pre-measure and post-measures in the intervention condition's average complexity level of predictions is considered unconvincing due to its small effect size ($M_{pre} = 4.13$ vs. $M_{post} = 4.03$, $p = 0.03$, $d = 0.2$). No change was found between the pre-measure and post-measure in the control condition's average complexity level ($M_{pre} = 4.09$ vs. $M_{post} = 4.05$, $p = 0.25$, $d = 0.12$). Similar results were found for the explanations, in that no change was found between pre-measure and post-measure in the intervention condition's average level of explanations ($M_{pre} = 4.21$ vs. $M_{post} = 4.26$, $p = 0.33$, $d = 0.05$) or in the control condition's average explanation level ($M_{pre} = 4.32$ vs. $M_{post} = 4.2$, $p = 0.18$, $d = 0.12$). As the analysis shows, there is a high probability that the differences are purely based on chance, which indicates no change in the intervention condition for the average level of complexity.

Conclusion and Discussion

The results of the present study indicate that professional development focused on stimulating thought-provoking teacher-student interactions has a positive impact on students' display of scientific knowledge. We outlined that there are different types of knowledge to be gained during science and technology lessons. With regard to the standardized measure, no difference was found between the intervention and control conditions. With regard to the situated measures, the students in the intervention condition improved between pre-measure to post-measure in two out of three variables (proportion and number of predictions and explanations), while no change was found in the control condition. No difference between the conditions was found with regard to the average level. This indicates that the classes who participated in the VFCt profited from the intervention. During the intervention, the proportion of more complex explanations and predictions steadily increased, indicative of more active participation of the students. Active participation of students in meaningful learning processes is considered to be important to better retain the course material and to become self-regulated learners. Educational psychologists, therefore, stress the importance of transferring the responsibility to learn to students instead of transferring knowledge to them (e.g., Kagan 2004; Ryan and Deci. 2000).

In conclusion, the results show that this intervention, by focusing on the actual, situated thinking process in authentic situations, led to meaningful changes in students' situated knowledge. An increase in the number of classroom interactions is important, as it implies an increase in the active construction of knowledge (Topping and Trickey 2014). Newton and Newton (2000) stressed that students' understanding can benefit from the types of questions teachers ask, but that it also depends on both the quality and quantity of their own speaking. An increase in the number of complex utterances might signify a change in the traditional balance of classroom interaction (mostly teacher-dominated), in that students participate more actively during the lessons. In addition, our finding that there was an increase in the number of students' higher-order thinking skills (predicting and explaining) implies that students have more active cognitive engagement and that students are encouraged to take responsibility for their own learning and thinking processes and thereby to go beyond fact reproduction. As no differences over time or between conditions were found in the standardized assessment and average level of complexity, this study suggests that only focusing on standardized measurements or average scores might lead to distorted conclusions about the effectiveness of the intervention and that teachers who rely on the test score might be underestimating the potential of students.

Standardized Assessment Versus Situated Scientific Knowledge Assessment

One of the main messages that we want to convey in this paper is that in order to assess the effects of an intervention such as the VFCt, several methodologies should be used, in an attempt to find converging evidence on what the nature of the effects of the intervention could be. We argued that it is important that researchers explore the naturalistic situation in which the intervention takes place. Therefore, in this paper, rather than standardized assessment, situated assessment was used. Thereby, rich information was gathered by combining group level perspectives and longitudinal measurements over time.

The following considerations should be taken into account with regard to the differences in findings between standardized assessment and situated assessment. Firstly, the main focus of the standardized assessment was on reproduction of declarative knowledge, which is domain specific. Domain-specific knowledge may not neatly reflect the topics covered during science and technology lessons. In contrast, the skills gained during the science and technology lessons (scientific understanding) are considered to be domain general. One might therefore conclude that the paper-and-pencil test relates to learning processes—domain-specific knowledge reproduction—other than the processes by which students learned the relevant knowledge and skills—domain-general, socially situated reasoning (Bao et al. 2009; for similar results in the lower grades, see Van der Steen 2014).

Secondly, students' learning is typically of a situated nature, where performance emerges from the interaction between a student and teacher. During standardized assessment, students are not given any support and thus need to rely on their previously stored knowledge. The focus is on what students can do without any help. However, this, i.e., knowledge in isolation, is not how skills and knowledge are usually taught or used in real life (Crouch et al. 2007; Van Geert 1994). This implies the importance of taking into account how learning typically occurs, namely, in the context of actual reasoning, instead of only taking into account the context of individual test performance. The interpretation of standardized assessments should therefore be done cautiously, and preferably in combination with contextual information (Koretz 2008). This study seems to indicate that only focusing on a single test score or average level might lead to underestimation of the dynamically situated and learning-oriented features of the cognitive skills.

A third explanation, suggested by Dejonckheere et al. (2009), is that the effect of improved scientific reasoning skills may not be directly visible in increased declarative knowledge levels, because the skills tend to have an indirect effect. The indirect effect refers to the assumption that scientific understanding provides students with opportunities to critically reflect upon experiences and to be inquisitive toward the world around them, which will eventually lead to expansion of their knowledge base. These skills are considered essential for full participation in society but appear hard to capture in a paper-and-pencil test.

Average Versus Number of Situated Scientific Knowledge

Although an increase was found in the number of utterances representing higher complexity levels (4, 5, 6, 7), this increase in frequency did not coincide with an increase in the average complexity level. This divergence in findings might be explained by a huge increase in the single representational level, which skews the findings (averages). Note that this average level is not a reflection of the optimal level students have reached during the lesson (see Fischer 1980).

However, the finding of no changes in the average complexity level of both explanations and predictions stands in contrast to the previous findings of pilot studies in the upper grades (Van Vondel et al. 2016) and the lower grades (Wetzels 2015) of primary education. This contradiction may be accounted for by some differences between this study and the previous studies. Firstly, compared to a pilot study (Van Vondel et al. 2016), the difference may be explained by how the data was collected. In the pilot study, the teachers were instructed to focus on air pressure and to aim at teaching students about high and low pressure both during the pre-measures *and* the post-measures. The similar content of the lessons might have led to a more equal comparison, as the average level of thought is highly dependent on the nature of the scientific content being taught (Meindertsma 2014). In some scientific domains, there is much to be gained in terms of average level of complexity, whereas in others, there is much more to be gained in terms of the number of questions asked, the number of hypotheses and explanations formulated, and so forth. In this study, teachers were free to select both the topic, as long as it fell within the general theme of the earth and space system, and the type of lesson of their choice. Although the coding scale used is considered to be task independent (Meindertsma 2014), a similar topic may well make a better comparison between lessons possible. Secondly, another explanation might be found in the way the average level was computed, i.e., which data it was based on. For instance, while Wetzels et al. based their results upon the highest complexity score during each minute of the lesson, the current study took all complex utterances of students into account, as each utterance is a reflection of the level of scientific understanding at that specific moment.

Limitation

The results of the present study are promising in that changes were found in students' scientific knowledge in the intervention condition. However, they should be interpreted in light of some potential procedural and methodological limitations. Generalization of the results should be done cautiously as the study took place with the help of a small sample of teachers, whose recruitment and assignment to one of the conditions was not carried out randomly. The choice of sample size and the selection of teachers for the experimental condition can be underpinned as follows. Firstly, it is important to note that the use of real-time authentic measures of scientific knowledge as a progress criterion is very labor intensive. This implies that a regular-sized study is almost automatically limited to a relatively small sample of teachers. Secondly, it is important to start effectiveness research with a selection of motivated participants (Ryan and Deci. 2000). If the intervention turns out to be effective with motivated participants, one can eventually scale it up in order to reach the entire population of teachers. Though these are limitations of the study, they are not trivial limitations in the sense that they could have been overcome if the researcher had spent more effort and thought on the design of the intervention study.

Future Studies

This study provides suggestions for several possible directions for future investigations. Firstly, though we use the term *situated* in this study, the analysis can be done much more dynamically, for example, by focusing on the unfolding of individual teaching-learning pathways over time, using microgenetic, time-serial measures. This study adds to previous findings by focusing on how changes in the classrooms over the course of the intervention emerged out of the microdynamics of interaction between teacher and students (i.e., the dynamics of real-time interaction between teacher and students).

Secondly, a follow-up might provide insight into whether or not the increased scientific understanding results in a long-term change in the knowledge level of students, for instance, at the end of the academic year, and whether or not the skills needed for scientific understanding become part of students' standard way of gaining knowledge (i.e., whether the newly learned skills are internalized). Several studies stress that when students gain insight into their own thinking processes and become increasingly capable of reflecting on problems, this reinforces the process of knowledge development (Christie et al. 2009; Dejonckheere et al. 2009).

Thirdly, a consideration for further study is to focus on students' level of engagement during actual science and technology activities (Laevers 2005). Several studies stress the importance of supporting students' interest and motivation, as expressed in the level of engagement, to attain higher academic achievement (e.g., Pietarinen et al. 2014). These underlying dynamics might provide a more complete understanding of the process of how teachers teach students particular ways of reasoning and of what elements are beneficial for students' engagement and the construction of scientific knowledge. These findings could then be used to strengthen evidence-based practice.

Implications on Theory and Practice

On a more general level, the results of this study may be tangent to a general view on assessing educational quality. Educational quality is often assessed focusing on teacher characteristics and student learning outcomes (Gorard 2013). This may be so because teachers (and policy makers) often want to know the effects of innovative programs on the kinds of group-administered standardized tests they are usually held accountable for. Basically, if the situated changes are not reflected in standardized measures, this might imply that students are often assessed in ways that do not optimally reflect the change in their skills. The fact that learning gains emerge in transaction and that the quality of student-teacher interaction in schools is linked to improved learning is often overlooked (Barber and Moursched 2007; Gorard 2013). However, in daily practice, the knowledge and insights constructed in transaction with the context is what students need and use. This actual change in behavior is essential for teachers as this accurately reflects what they actually do and what they experience during their lessons. Information about actual changes in behavior provides deeper insights into how teachers can optimize their lessons—compared to standard evaluations.

The fact that the standardized and situated assessments offer alternative conclusions does not invalidate either of them, but rather shows that different methods of analysis expose different layers of classroom practices. However, the point is that students are typically only assessed using standardized measures in educational settings and that effects of interventions are assessed in such a way as well. We argue that the desired goal of science and technology education is not factual science and technology knowledge, but understanding of the process of doing science. If such 21st century skills need to be taught and assessed, a shift in teaching and assessment seems important. Factual scientific knowledge is constantly being updated and expanded through new technological highlights and increased insights. The current theory is that human knowledge is doubling every 13 months (internetworldstats.com). Transferring such knowledge to students would be an impossible task. A much more appropriate endeavor is therefore to teach students how to approach the world of science. Students should be encouraged to become self-regulated learners and stimulated to critically approach natural phenomena (instead of simply accepting "evidence"). An important learning outcome is then that students use reasoning skills to connect the situation at hand with previously acquired knowledge. Teachers can stimulate (and assess) students' progress during the activities by explicitly evoking their current level of understanding during a specific task and support the students to extend the range of their level of complexity. Overall, for teachers there is no one correct way to teach science and technology. However, some elements of the VFCt that proved important are using pedagogical-didactical strategies such as using the steps of the empirical cycle as a guide to structure the experiments and thinking of students, using questions to evoke students' display of scientific knowledge and curiosity, and taking the role of a coach to assist and scaffold students in structuring their own learning processes. The most important aspect of effective science and technology teaching is to actively evoke students' display of scientific knowledge by means of open questions. A practical recommendation for teachers is then to occasionally review video observations of their own lessons as a means to increase their quality of instructional practice. Teachers' awareness of their own role is an important indicator for the quality of their support, which is considered a crucial factor in improving students' learning (Barber and Moursched 2007).

Concluding Remark

From a policy and practice perspective, we believe that the present study provides important information about an effective form of educational intervention and suggestions for how to assess the effectiveness of science-educational interventions. The results provided evidence that the use of video feedback coaching is an effective way to change teachers' practice and, by doing so, it proved an effective way for changing the classroom system towards patterns of interaction that promote students' display of situated scientific knowledge. It provides an argument for incorporating situated assessment of the evaluation of teachers' professional development in science education. Hence, it is important not only to focus on gains in scientific knowledge but also on gains in the types of support that can be given during classroom interaction. The importance of incorporating situated test situations in the assessment of students' progress is thereby stressed. This study supports the notion that it is important to examine students' development and the effects of an intervention that uses process measures (Van Geert 1994). These measures show insight into changes in situ and are important additions to questionnaire and test data.

Acknowledgments The research described was funded by the Curious Minds ("TalentenKracht") program; supported by the Dutch Ministry of Education, Culture, and Science and the National Platform Science and Technology (Platform Bètatechniek); and co-funded by the Heymans Institute (University of Groningen).

Compliance with Ethical Standards The teachers and parents of each participating student gave active consent before the start of the study. The study was approved by the Ethical Committee Psychology of the University of Groningen (the Netherlands).

References

Bao, L., Cai, T., Koenig, K., Fang, K., Han, J., Wang, J., et al. (2009). Physics learning and scientific reasoning. *Science, 323*(5914), 586–587. doi:10.1126/science.1167740.

Barber, M., & Moursched, M. (2007). *How the world's best-performing school systems come out on Top*. McKinsey and Co., London.

Bilgin, İ. (2006). The effects of hands-on activities incorporating a cooperative learning approach on eight grade students' science process skills and attitudes towards science. *Journal of Baltic Science Education, 1*(9), 27–37.

Blickenstaff, C. J. (2005). Women and science careers: leaky pipeline or gender filter? *Gender and Education, 17*(4), 369–386. doi:10.1080/09540250500145072.

Borman, G. D., Gamoran, A., & Bowdon, J. (2008). A randomized trial of teacher development in elementary science: first-year achievement effects. *Journal of Research on Educational Effectiveness, 1*, 237–264. doi:10.1080/19345740802328273.

Bos, J., & Steenbeek, H. (2009). *Mediacoder: software voor het coderen van video- en audio-materialen*. Groningen: Internal publication, IDP Department, University of Groningen.

Bybee, R. W., & Fuchs, B. (2006). Preparing the 21st century workforce: a new reform in science and technology education. *Journal of Research in Science Teaching, 43*(4), 349–352. doi:10.1002/tea.20147.

Chin, C. (2006). Classroom interaction in science: teacher questioning and feedback to students' responses. *International Journal of Science Education, 28*(11), 1315–1346. doi:10.1080/09500690600621100.

Christie, D., Tolmie, A., Thurston, A., Howe, C., & Topping, K. (2009). Supporting group work in Scottish primary classrooms: improving the quality of collaborative dialogue. *Cambridge Journal of Education, 39*(1), 141–156. doi:10.1080/03057640802702000.

Cito [Dutch National Institute for Measurement in Education] (2010). *Wereldoriëntatie deel 1: Natuuronderwijs [World orientation part 1: Nature education]*. Arnhem: Cito.

Crouch, C. H., Watkins, J., Fagen, A. P., & Mazur, E. (2007). Peer instruction: engaging students oneon-one, all at once. *Research-Based Reform of University Physics 1*(1), 40–95.

Cuevas, P., Lee, O., Hart, J., & Deaktor, R. (2005). Improving science inquiry with elementary students of diverse backgrounds. *Journal of Research in Science Teaching, 42*(3), 337–357. doi:10.1002/tea.20053.

De Groot, A. D. (1994). *Methodologie: grondslagen van onderzoek en denken in de gedragswetenschappen. ‘s*. Gravenhage: Mouton and Co.

Dejonckheere, P. J. N., Van De Keere, K., & Mestdagh, N. (2009). Training the scientific thinking circle in pre- and primary school children. *The Journal of Educational Research, 103*(1), 1–16. doi:10.1080/00220670903228595.

Dekker, B., Krooneman, P. J., & Van der Wel, J. J. (2008). *Bronnenboek VTB: stand van zaken 2007–2008.Technisch rapport [Source book for the expansion of technology in primary education: progress report 2007–2008. Technical report]*. Amsterdam: Regioplan.

Durlak, J. A., & DuPre, E. P. (2008). Implementation matters: a review of research on the influence of implementation on program outcomes and the factors affecting implementation. *American Journal of Community Psychology, 41*(3–4), 327–350. doi:10.1007/s10464-008-9165-0.

Esmeijer, J., & Van der Plas, A. (2012). *Innovatiekaart : empowered learning in the 21st century*. Delft: TNO.

Fischer, K. W. (1980). A theory of cognitive development: the control and construction of hierarchies of skills. *Psychological Review, 87*, 477–531. doi:10.1037/0033-295X.87.6.477.

Fischer, K. W., & Bidell, T. R. (2006). Dynamic development of action, thought, and emotion. In R. M. Lerner (Ed.), *Handbook of child psychology. Vol 1: theoretical models of human development* (6th ed., pp. 313–399). New York: Wiley.

Fischer, K. W., & van Geert, P. L. C. (2014). Dynamic development of brain and behavior. In P. Molenaar, R. Lerner, & K. Newell (Eds.), *Handbook of developmental systems theory and methodology* (pp. 287–314). New York: Guilford Press.

Fouad, N. a., Hackett, G., Smith, P. L., Kantamneni, N., Fitzpatrick, M., Haag, S., & Spencer, D. (2010). Barriers and supports for continuing in mathematics and science: gender and educational level differences. *Journal of Vocational Behavior, 77*(3), 361–373. doi:10.1016/j.jvb.2010.06.004.

Fredrickson, B. L. (2015). The broaden-and-build theory of positive emotions. *Philosophical Transactions, 359*(1449), 1367–1377. doi:10.1098/rstb.2004.1512.

Fukkink, R. G., Trienekens, N., & Kramer, L. J. C. (2011). Video feedback in education and training: putting learning in the picture. *Educational Psychology Review, 23*(1), 45–63. doi:10.1007/s10648-010-9144-5.

Furtak, E. M., Seidel, T., Iverson, H., & Briggs, D. C. (2012). Experimental and quasi-experimental studies of inquiry-based science teaching: a meta-analysis. *Review of Educational Research, 82*(3), 300–329. doi:10.3102/0034654312457206.

Gibson, H. L., & Chase, C. (2002). Longitudinal impact of an inquiry-based science program on middle school students' attitudes toward science. *Science Education, 86*(5), 693–705. doi:10.1002/sce.10039.

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics, 33*, 587–606. doi:10.1016/ j.socec.2004.09.033.

Gorard, S. (2013). What difference do teachers make? A consideration of the wider outcomes of schooling. *Irish Educational Studies 32*(1), 69–82.

Hargreaves, L., Moyles, J., Merry, R., Paterson, F., Pell, A., & Esarte-Sarries, V. (2003). How do primary school teachers define and implement “interactive teaching” in the national literacy strategy in England? *Research Papers in Education, 18*(3), 217–236. doi:10.1080/0267152032000107301.

Hintze, J., & Matthews, W. J. (2004). The generalizability of systematic direct observations across time and setting: a preliminary investigation of the psychometrics of behavioral observation. *School Psychology Review, 33*(2), 258–270.

Hood, G. (2004). *PopTools version 3.2.5*. Retrieved November 11, 2014, from http://www.poptools.org/

Jorde, D., & Dillon, J. (2013). *Science education research and practice in Europe: retrospective and prospective*. The Netherlands, Rotterdam: Sense Publishers.

Kagan, S. (2004). *From lessons to structures – A paradigm shift for 21st century education*. Kagan Publishing, San Clemente, CA.

Kennedy, H., Landor, M., & Todd, L. (Eds.) (2011). *Video interaction guidance: a relationship-based intervention to promote attunement, empathy and wellbeing*. London: Jessica Kingsley Publishers.

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

Koretz, D. M. (2008). *Measuring up*. United States: Harvard University Press.

Kuhn, D. (2010). Teaching and learning science as argument. *Science Education, 94*(5), 810–824. doi:10.1002/sce.20395.

Laevers, F. (2005). Ervaringsgericht onderwijs als antwoord. *EE-M@gazine,* 1–3.

Lee, Y., & Kinzie, M. B. (2011). Teacher question and student response with regard to cognition and language use. *Instructional Science, 40*(6), 857–874. doi:10.1007/s11251-011-9193-2.

Lutz, S. L., Guthrie, J. T., & Davis, M. H. (2006). Scaffolding for engagement in elementary school reading instruction. *Journal of Educational Research, 100*(1), 3–20. doi:10.3200/JOER.100.1.3-20.

Meindertsma, H. B. (2014). Predictions and explanations: short-term processes of scientific understanding in young children (Dissertation). University of Groningen.

Miller, S. P., & Hudson, P. J. (2007). Using evidence-based practices to build mathematics competence related to conceptual, procedural, and declarative knowledge. *Learning Disabilities Research and Practice, 22*(1), 47–57. doi:10.1111/j.1540-5826.2007.00230.x.

Minner, D. D., Levy, A. J., & Century, J. (2010). Inquiry-based science instruction—what is it and does it matter? Results from a research synthesis years 1984 to 2002. *Journal of Research in Science Teaching, 47*(4), 474–496. doi:10.1002/tea.20347.

Mortenson, B. P., & Witt, J. C. (1998). The use of weekly performance feedback to increase teacher implementation of a prereferral academic intervention. *School Psychology Review, 27*(4), 613.

Mullis, I.V.S., Martin, M.O., Ruddock, G.J., O'Sullivan, C.Y., and Preuschoff, C. (2011). *TIMSS 2011 Assessment Frameworks*. Boston, MA: TIMSS and PIRLS International Study Center Lynch School of Education, Boston College.

Newton, D. P., & Newton, L. D. (2000). Do teachers support causal understanding through their discourse when teaching primary science? *British Educational Research Journal 26*(5), 599–613. doi:10.1080/713651580.

NGSS Lead States (2013). *Next generation science standards: for states, by states*. Washington, DC: The National Academies Press.

Noell, G. H., Witt, J. C., Slider, N. J., Connell, J. E., Gatti, S. L., Williams, K. L., et al. (2005). Treatment implementation following behavioral consultation in schools: a comparison of three follow-up strategies. *School Psychology Review, 34*(1), 87–106.

Oliveira, A. W. (2009). Developing elementary teachers' understanding of the discourse structure of inquiry-based science classrooms. *International Journal of Science and Mathematics Education, 8*, 247–269.

Oliveira, A. W. (2010). Improving teacher questioning in science inquiry discussions through professional development. *Journal of Research in Science Teaching, 47*(4), 422–453. doi:10.1002/tea.20345.

Penuel, W. R., Gallagher, L. P., & Moorthy, S. (2011). Preparing teachers to design sequences of instruction in earth systems science: a comparison of three professional development programs. *American Educational Research Journal, 48*(4), 996–1025. doi:10.3102/0002831211410864.

Pietarinen, J., Soini, T., & Pyhältö, K. (2014). Students' emotional and cognitive engagement as the determinants of well-being and achievement in school. *International Journal of Educational Research, 67*, 40–51. doi:10.1016/j.ijer.2014.05.001.

Rappolt-schlichtmann, G., Tenenbaum, H. R., Koepke, M. F., & Fischer, K. W. (2007). Transient and robust knowledge: contextual support and the dynamics of children's reasoning about density. *Mind Brain and Education, 1*(2), 98–108. doi:10.1111/j.1751-228X.2007.00010.x.

Reinke, W. M., Sprick, R., & Knight, J. (2009). Coaching classroom management. In D. Alpert (Ed.), *Coaching: approaches and perspectives* (pp. 91–112). United States: Corwin Press.

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist 55*, 68–78. doi:10.1037/0003-066x.55.1.68.

Scardamalia, M., & Bereiter, C. (2006). Knowledge building: theory, pedagogy, and technology. In K. Sawyer (Ed.), *The Cambridge handbook of the learning sciences* (pp. 97–118). NY: Cambridge University Press.

Schneider, J. W. (2015). Null hypothesis significance tests. A mix-up of two different theories: the basis for widespread confusion and numerous misinterpretations. *Journal Scientometrics, 102*, 411–432. doi:10.1007/s11192-014-1251-5.

Schwartz, M. S., & Fischer, K. W. (2004). Building general knowledge and skill: cognition and microdevelopment in science learning. In A. Demetriou & A. Raftopoulos (Eds.), *Emergence and transformation in the mind: modeling and measuring cognitive change* (pp. 157 – 185). Cambridge, UK: Cambridge University Press.

Seidel, T., Stürmer, K., Blomberg, G., Koberg, M., & Schwindt, K. (2011). Teacher learning from analysis of videotaped classroom situations: does it make a difference whether teachers observe their own teaching or that of others? *Teaching and Teacher Education, 27*(2), 259–267. doi:10.1016/j.tate.2010.08.009.

Silva, E. (2009). Measuring skills for 21st-century learning. *The Phi Delta Kappan, 90*(9), 630–634.

Şimşek, P., & Kabapinar, F. (2010). The effects of inquiry-based learning on elementary students' conceptual understanding of matter, scientific process skills and science attitudes. *Procedia – Social and Behavioral Sciences, 2*(2), 1190–1194. doi:10.1016/j.sbspro.2010.03.170.

Sullivan, G. M., & Feinn, R. (2012). Using effect size -- or why the *p* value is not enough. *Journal of Graduate Medical Education, 4*, 279–282. doi:10.4300/JGME-D-12-00156.1.

Tang, X., Coffey, J. E., Elby, A., & Levin, D. M. (2009). The scientific method and scientific inquiry: tensions in teaching and learning. *Science Education, 94*(1), 29–47. doi:10.1002/sce.20366.

Thelen, E., & Smith, L. B. (1994). *A dynamic systems approach to the development of cognition and action*. Cambridge: MA: MIT Press.

Thurston, A., Christie, D., Howe, C. J., Tolmie, A., & Topping, K. J. (2008). Effects of continuing professional development on group work practices in Scottish primary schools. *Journal of In-service Education, 34*(3), 263–282. doi:10.1080/13674580802264803.

Todman, J. B., & Dugard, P. (2001). *Single–case and small–n experimental designs: a practical guide to randomization tests*. Mahwah (NJ): Erlbaum.

Topping, K. J., & Trickey, S. (2014). The role of dialog in philosophy for children. International *Journal of Educational Research 63*, 69–78. doi:10.1016/j.ijer.2013.01.002.

Van de Pol, J., Volman, M., & Beishuizen, J. (2010). Scaffolding in teacher-student interaction: a decade of research. *Educational Psychology Review 22*(3), 271–296. doi:10.1007/s10648-010-9127-6.

Van de Pol, J., Volman, M., & Beishuizen, J. (2011). Patterns of contingent teaching in teacher–student interaction. *Learning and Instruction, 21*(1), 46–57. doi:10.1016/j.learninstruc.2009.10.004.

Van der Steen, S. (2014). "How does it work?": a longitudinal microgenetic study on the development of young children's understanding of scientific concepts (Dissertation). University of Groningen.

Van der Steen, S., Steenbeek, H. W., Van Dijk, M. W. G., & Van Geert, P. L. C. (2014). A complexity approach to student's understanding of scientific concepts: a longitudinal case study. *Learning and Individual Differences, 30*, 8–91. doi:10.1016/j.lindif.2013.12.004.

Van der Steen, S., Steenbeek, H. W., & Van Geert, P. L. C. (2012). Using the dynamics of a person-context system to describe student's understanding of air pressure. In H. Kloos, B. J. Morris, & J. L. Amaral (Eds.), *Current topics in Student's learning and cognition* (pp. 21–44). Rijeka, Croatia: InTech. doi:10.5772/53935.

Van Geert, P. L. C. (1994). *Dynamic systems of development: change between complexity and chaos*. NY: Harvester Wheatsheaf.

Van Keulen, H., & Sol, Y. (Eds.) (2012). *Talent ontwikkelen met wetenschap en techniek*. Utrecht: Onderwijsadvies en Training, Centrum voor Onderwijs en Leren.

Van Vondel, S., Steenbeek, H. W., Van Dijk, M. W. G., & Van Geert, P. L. C. (2016). 'Looking at' educational interventions. Surplus value of a complex dynamic systems approach to study the effectiveness of a science and technology educational intervention. In M. Koopmans & D. Stamovlasis (Eds.), *Complex dynamical systems in education: concepts, methods and applications* (pp. 203–232). New York: Springer. doi:10.1007/978-3-319-27577-2.

Van Zee, E., Minstrell, J., & Minstrel, J. (1997). Using questioning to guide student thinking. *Journal of the Learning*, 37–41. doi:10.1207/s15327809jls0602.

Wade, R. K. (1984). What makes a difference in inservice teacher education? A meta-analysis of research. *Educational Leadership, 42*(4), 48.

Webb, N. M. (1989). Peer interaction and learning in small groups. *International Journal of Educational Research, 13*, 21–40. doi:10.1016/0883-0355(89)90014-1.

Wetzels, A. (2015). Curious minds in the classroom: the influence of video feedback coaching for teachers in science and technology lessons (Dissertation). University of Groningen.

White, B. Y., & Frederiksen, J. R. (1998). Inquiry, modeling, and metacognition: making science accessible to all students. *Cognition and Instruction, 16*(1), 3–118. doi:10.1207/s1532690xci1601_2.

Zimmerman, C. (2000). The development of scientific reasoning skills. *Developmental Review, 20*(1), 99–149. doi:10.1006/drev.1999.0497.