Curious minds in the classroom

Wetzels, Anna

Publication date:
2015

[Link to publication in University of Groningen/UMCG research database](#)

CHAPTER 5:

Primary Science Teaching: Behavior of Teachers and their Pupils during and after a Coaching Program

CHAPTER 5:
Primary Science Teaching: Behavior of Teachers and their Pupils during and after a Coaching Program

5.1 Introduction

5.1.1 The importance of science in school

As the OECD (2008) states: " In today's technology-based societies, understanding fundamental scientific concepts and theories and the ability to structure and solve scientific problems are more important than ever" (p. 16). In current society, public knowledge of science is important for various reasons. First, it is important because science, as such, underlies the technology that everybody is dependent upon. Second, knowledge of science is important because it underlies peoples' decisions regarding socio-scientific issues such as gene manipulation and climate change. Finally, knowledge of science is important because it strengthens the critical attitude of the public towards phenomena like astrology or quack doctors.

These examples show that science is essential for the way society operates as a complex system, now and in the years to come (Nowotny et al., 2001). The implication thereof is that throughout society more science and science related knowledge is needed (Mooney & Kirshenbaum, 2009). Science education from the earliest school years on is an important means to achieve the goal to obtain this knowledge.

Therefore, it is important to stimulate children's science learning potential from preschool on. This potential expresses itself in children at a young age who show that they are already little scientists (Brewer, Chinn, & Samarapungavan, 1998), and have a great interest in science topics (Chouinard, 2007). However, when these children go to school, the great majority of them lose much of this natural talent for science learning (Engel, 2009). That is, children lose their interest in science, their curiosity, as well as their natural way of asking questions and reasoning. As enthusiasm for, and interest in, science can be lost during these early school years, it should be possible for teachers to stimulate these features in young children during school hours. In addition, parents can stimulate this enthusiasm and interest at home, by encouraging their children to express and satisfy their curiosity (Tizard & Hughes, 2002). However, science has not been given much attention in early childhood classrooms (Appleton, 2003; Dickinson et al., 1997; Martin, Mullis, & Foy, 2008; Michaels, Shouse, & Heidi, 2007). In order to help children from a young age on to develop the much needed science

knowledge, together with the related scientific reasoning skills, schools should give more attention to science learning from preschool on. In their science lessons, they should focus not only on teaching science content and enabling pupils to learn this content, but also on stimulating other aspects of their pupils' scientific learning behavior, such as their interest in science, their enthusiasm for science, and their expression of scientific reasoning skills, such as those needed in predicting and explaining scientific phenomena.

A common way of teaching science in schools is by means of inquiry learning (Benford & Lawson, 2001; Furtak, 2006; Van Graft & Kemmers, 2007; "What is Enquiry-Based Learning (EBL)?," n.d.). Pupils' inquiry learning is aimed at learning how to think scientifically and independently. According to Dewey (1997), inquiry learning is 'a way of learning science as a process and a way of thinking', which means that it is not primarily focused on science content, i.e., on the knowledge of facts, but on the methods and processes scientists use. It can be used in science lessons to help both teachers and pupils to teach and learn science in interaction with each other. Doing experiments, and predicting and explaining what happens during these experiments (Van Joolingen, De Jong, & Dimitrakopoulou, 2007), helps pupils to think critically and to ask scientific questions. This can then be used to enable them to make sense of their observations in the natural world. Thus, inquiry learning can help to stimulate pupils' scientific reasoning skills. Inquiry learning demands other teaching-learning behavior from both teachers and pupils than what is used in mathematics or English lessons. With inquiry learning, teachers need to learn how to provide adequate support for stimulating pupils' scientific reasoning skills, which requires teachers to behave in a way that they frequently are not accustomed to. This support is visible in the interaction between teacher and pupil, in which pupils can develop the scientific reasoning skills needed for inquiry learning.

Teachers are often not educated in science and inquiry learning, and often lack knowledge of science themselves. As teachers are role models for children, it should be self evident that their behavior in the classroom regarding science education has a clear impact on young children's inquiry skills. An intervention focused not only on teachers' science knowledge, but also on their behavior regarding science lessons in the earliest school years, should therefore be the starting point for enhancing science education in schools.

The aim of this chapter is to study whether a video-feedback coaching intervention for earliest school year teachers can contribute to teachers' behavioral change with respect to stimulating pupils' scientific reasoning skills, and subsequently to the level of pupils' scientific reasoning skills themselves. In this specific

form of a professional development trajectory, teachers learn scientific reasoning skills themselves as a way to improve their science teaching behavior, so they can be good role models for their pupils by showing their knowledge for, and interest in, science and scientific reasoning. A multiple case study, using micro genetic measures with regard to change in teachers' and pupils' behavior and their interaction patterns in the classroom during science lessons, is carried out to explore the effects of the intervention on both teachers and pupils.

5.1.2 Knowledge needed when teaching science

Overall, science class in the earliest grades is provided by a teacher, who has not been trained as a specialized science teacher. These teachers usually do not have any experience in teaching science, and do not know enough about science themselves. In addition, they do not know how science can best be taught, and, as a result, they do not feel very confident about their abilities to teach science (Dickinson et al., 1997). Moreover, they often teach science in an authoritarian way or in a way that does not spark enthusiasm and interest in science, i.e., merely using textbooks or worksheets (Dickinson et al., 1997).

The three important types of knowledge needed for teaching science as described in the educational literature (Smith, 1999) are general pedagogical knowledge, science content knowledge, and pedagogical content knowledge, i.e., the way science knowledge is used in classes with the intention to understand and impose meaning and thus the way science can be taught (Hattie, 2003; Smith, 1999).

The first type of knowledge, general pedagogical knowledge, is needed for classroom organization and management, instructional models and classroom discourse (Morine-Dershimer & Kent, 1999; Shulman, 1987). Classroom organization and management stresses the importance of content based instruction, active learning for pupils, as well as the provision of learning activities in line with the pupils' level of thinking (Brophy, 1999; Brophy & Good, 1986). Instructional models include the use of an adequate mix of direct instruction and scaffolding, depending upon the amount of structure a task has for students (Rosenshine & Stevens, 1986). The use of classroom discourse, particularly discourse which uses experience based questions, questions requiring observations and analyses of available data, is known to yield higher student achievements (Otto & Schuck, 1983).

In addition to the general pedagogical knowledge, teachers need to obtain adequate science content knowledge, i.e., an understanding of the basic scientific principles at the level of their pupils, as well as pedagogical content knowledge at that same level. Primary school teachers often claim they lack science content knowledge (Appleton, 2003; Dickinson et al., 1997; Marx & Harris, 2006). There

is, however, some discussion in the literature about content knowledge. On the one hand, science content knowledge is deemed necessary to teach science in a proper way; on the other hand, there is a group of researchers who state that the importance of content knowledge (CK) is overrated (Zeidler, 2002), and that teachers need pedagogical content knowledge as a more specific body of knowledge (Ball, Lubienski, & Mewborn, 2001).

The third kind of knowledge is, in fact, a combination of the first and second type of knowledge, namely pedagogical content knowledge. This is knowledge of how to teach specific content in specific contexts; a form of knowledge in action (Mellado, Blanco, & Ruiz, 1998; Shulman, 1987), and consists partially of content knowledge and partially of pedagogical knowledge. It includes teachers' orientation to teaching science, knowledge of science curricula, knowledge of assessment, knowledge of scientific literacy, knowledge of students' understanding of science, knowledge of science content, and knowledge of instructional strategies (Magnusson, Krajcik, & Borko, 1999). In addition to, and partially because of a lack of science content knowledge, teachers do not possess enough pedagogical content knowledge in science, nor the skills to put this knowledge into practice. Because of their lack of content knowledge and pedagogical content knowledge when teaching science, teachers use ineffective teaching methods, such as teacher discussions and explanation, watching science television shows, library results, demonstrations and work sheets (Appleton, 2003; Goodrum et al., 2001), all of which fail to give pupils sufficient opportunities for reasoning scientifically. Teachers need to learn and have experience with effective teaching methods, such as content based instruction, active learning for pupils, as well as providing learning activities in line with pupils' level of thinking (Brophy, 1999; Brophy & Good, 1986) and classroom discourse with the use of thought provoking questions as these are all activities that enhance pupils' scientific reasoning skills.

Pure science content knowledge, therefore, is not fundamental in the science lessons for young children. However, for young pupils' learning to be effective, pedagogical science content knowledge is necessary when focusing on the processes and methods used by scientists and technicians (Van Joolingen et al., 2007) by using inquiry learning. For this reason, teachers need to have insight into the instructional strategies that can be used within inquiry based science education.

5.1.3 Instructional strategies when teaching science

Instructional strategies are techniques teachers use to help pupils become independent, strategic learners. Instructional strategies have - amongst other things - the aim of motivating pupils, of focusing their attention, and of organizing information for understanding and remembering ("Instructional Strategies," 2002). In

this article, we focus on three important instructional strategies that comprise a mixture of science content and pedagogical content knowledge, and that together offer teachers clear guidance for implementing inquiry based science teaching. The first strategy is the use of the empirical cycle (De Groot, 1961), which gives teachers and children structure during inquiry learning; the second strategy is asking questions, which provides teachers with a tool to give pupils more room for scientific reasoning; and the third strategy is scaffolding, which gives teachers a good means for supporting pupils' scientific reasoning skills. In the remainder of this section, these three strategies will be discussed.

In his book, 'Methodology: foundations of inference and research in the behavioral sciences', De Groot (1961) describes the empirical cycle as an often used strategy in science and scientific research. De Groot's model includes several important learning activities, such as hypothesizing, observing, explaining and reasoning. The empirical cycles supports inquiry learning by using the following steps: conducting the draft of a research question, formulating a hypothesis concerning the phenomena that are being studied, setting up an experiment to demonstrate the truth or falseness of the hypothesis, observing what happens during the experiment, and finally, drawing a conclusion that validates or modifies the hypothesis. From this last step a new research question can emerge, implying that the cycle is repeated with step one (De Groot, 1961; Dejonckheere et al., 2009).

The second strategy, asking questions, stimulates pupils' curiosity (Goodman & Berntson, 2000), and gives them plenty of opportunities to learn (Wasik et al., 2006), to reason, and to explore (Lee, 2010). However, the questions teachers ask are usually focused on reproduction of knowledge and facts, instead of on enhancing thinking (Engel, 2009; Engelhard & Monsaas, 1988). Asking the right questions includes two main aspects. First, unlike closed questions, open questions do not limit pupils' answers (Hargreaves, 1998; Rivera et al., 2005), but instead elicit more elaborated answers. Secondly, pupil centered questions (e.g. "what do you think what will happen if I pour oil in water") give room to pupils' thinking and reasoning, in contrast to teacher oriented questions, which ask for reproduction of knowledge (e.g. "what is an atom?"(Oliveira, 2010).

The third strategy, scaffolding, gives teachers the opportunity to help pupils until it is no longer needed, and gives pupils the opportunity to work independently, and to receive help only to the extent that it enables them to again work on their own (Mayer, 2004; Palincsar & Brown, 1984). Pupils have a cognitive difference between what they show that they know and what they are potentially capable of (the zone of proximal development (Chaiklin, 2003; Vygotsky, 1978)). This zone

can be reached by using scaffolding, i.e., customized help during a task of a pupil who has a higher level of cognitive abilities. This help enables pupils to accomplish tasks they could otherwise not complete themselves.

All three strategies independently have a positive effect on children's scientific reasoning skills. Combining these strategies in a professional development trajectory for teachers aimed at teaching science during the first school years seems, therefore, to be a logical step. In the current study, the three strategies are combined.

5.1.4 Teachers' professional development

The knowledge and skills teachers need in order for them to actually change their behavior should be obtained in a professional development trajectory. That is, knowledge alone is not sufficient in that it will not change behavior in the classroom (Birman et al., 2000). Teachers need time to digest and practice new knowledge. Several studies have shown that there are more elements needed to yield a lasting behavioral change (Desimone, 2009; Garet et al., 2001). The elements needed include sufficient time to learn the new way of working, coherence of the learning activity with their daily work and school policy, focus on the learning process of children, active learning and room to reflect on the learning process, and finally, learning together with teachers from the same school (Birman et al., 2000; Desimone, 2009).

However, a professional development program that focuses on learning new skills and knowledge alone will in general not be effective (Han & Weiss, 2005). Attention must be paid to teachers' intrinsic motivation to change their behavior and to start using the newly learned strategies in the classroom. To enhance intrinsic motivation, three basic concerns that all human beings have, namely the concerns for competence, autonomy and relatedness, need to be addressed (Minnaert et al., 2007; Ryan & Deci, 2000; Steenbeek & Van Geert, 2007, 2013). When all three concerns are adequately met, teachers work intrinsically motivated, and show more enthusiasm, which leads to a higher quality of teaching in the classroom (Kunter et al., 2008).

A professional development trajectory must, therefore, include knowledge about general teaching strategies, the opportunity to practice the learned content and a way to address motivation issues. Such a trajectory in the form of a coaching program can provide teachers with tools to reach behavioral change, and thereby a higher quality of teaching, which in turn will have a positive effect on pupils' scientific reasoning skills. The effects of such a coaching program can be measured in practice by counting the number of scientific reasoning eliciting

questions teachers ask and by calculating the level of scientific reasoning skills pupils achieve, as shown in their verbal utterances.

5.1.5 Current study

The study examines the effect of a video coaching program for elementary school teachers of 4-8 year old pupils, (Video Feedback Coaching for Teachers, (VFC-T; Wetzels et al. 2011), which is focused on behavioral change. The study comprises three elements, knowledge about the empirical cycle, asking questions and scaffolding, then, the opportunity to put this knowledge in practice, and, finally, attention to teachers' intrinsic motivation. In this study, we consider the effectiveness of the VFC-T on teachers, and subsequently on the scientific reasoning skills of their pupils, with regard to their classroom discourse, over the duration of the coaching program. Due to limitations of the coaching program and the design of the study, we restrict ourselves to reporting quantitative results regarding verbal utterances during the lessons, by focusing on teachers' use of scientific reasoning eliciting questions and pupils' use of remarks that reflect their scientific reasoning skills.

The study is framed by the following questions:

1. To what extent does the intervention influence teachers' behavior with regard to the use of questions that elicit scientific reasoning eliciting in the classroom?
2. To what extent does the intervention influence pupils' level of scientific reasoning in the classroom, as shown in their verbal utterances?
3. a. What patterns of change can be recognized in individual teacher's trajectories towards increasing the use of questions that elicit scientific reasoning?
   b. What level of coherence is there between the patterns of change in individual teachers' use of questions that elicit scientific reasoning and the level of scientific reasoning of their group of pupils?

During the program, we expect teachers to increase their number of questions that stimulate pupils' scientific reasoning, although the increase will differ for the various teachers (hypothesis 1a). We also expect the number of questions that stimulate pupils' scientific reasoning to increase in the experimental group and not in the control group (hypothesis 1b). Moreover, we expect pupils in the experimental group to show an increase in their level of scientific reasoning during the program (hypothesis 2a), and no increase in the control group (hypothesis 2b). Finally, we expect these results to have a sustainable effect, as will be shown in the post-measurement two months after the program ends (hypothesis 3). A further expectation is that inter-individual variability will occur, in that individual teachers will show different patterns of changes (hypothesis 4a). However, we do expect coherence with regard to the patterns of change visible in individual

teachers and their accompanying group of pupils. More specifically, we expect an increase over time in both variables (hypothesis 4b).

5.2 Method

5.2.1 Participants

The study was started after an extensive pilot study, in which the design of the study was developed and tested (Wetzels, Steenbeek, & Van Geert, 2015). Six elementary school teachers, all female, working with children aged 5-8 from two schools in the North of the Netherlands, took part in the intervention (mean age 51, range 35-61, mean experience 23 years, experience range 10-39 years). In addition, a group of five teachers of three schools from the same region, participated as a control group (mean age 39, range 22- 48 years old, mean experience 17 years, experience range 1-27). The teachers of the control group would have the opportunity to follow the intervention program in the next school year. The teachers of the experimental group and the control group were comparable in age and experience, except for one teacher in the control group. In all other aspects (such as experience in teaching science, voluntarily participation), the schools and the teachers were comparable. All participating schools, both experimental and controls, were interested in teaching science, which showed in their participation in "Beta Punt Noord", a partnership of primary education, one academic university, several universities of applied sciences and several businesses in the north of the Netherlands, with the aim of anchoring science and technology in the education program of primary schools. The teachers were asked to choose a group of four to six children, who would be representative for the class as a whole, to participate. All teachers and parents were informed about the use of video recordings, and all gave their written consent for using them for research.

5.2.2 Design

The study, including pre-intervention, coaching and post-intervention lessons, took place over a period of six months (from January until June). After two pre-intervention lessons, during which teachers taught a science lesson which they thought was suitable for their chosen small group of children, an introduction of two hours took place. This introduction was followed by four coaching sessions, during which the teachers taught a science lesson of their own choice, lasting twenty to thirty minutes, with the same group of pupils from their class. The time between the two coaching sessions ranged from one to two weeks. Immediately after each lesson, a coaching session of thirty to sixty minutes took place, using fragments that were selected from the video recording of the lesson of that day. Two months after the last coaching session, two post-intervention lessons were recorded in subsequent weeks. The intervention,

including the introduction and all coaching sessions, was performed using an extensive coaching manual (Wetzels et al., 2011) by the first author, who is a psychologist and a trained coach.

5.2.3 Content of the professional development trajectory for teachers

The aim of the coaching was to teach the teachers how to enhance their knowledge and skills of science teaching using instructional strategies as described above (section 1.3.), in order for the pupils to enhance their scientific reasoning skills during the science lessons. The coaching program is focused on teachers' behavioral change, and comprises three elements, knowledge about the empirical cycle, asking questions and scaffolding, then, the opportunity to put this knowledge into practice, and, finally, attention to teachers' intrinsic motivation. Firstly, during the introduction teachers received information about the following strategies: asking questions (Oliveira, 2010), scientific thinking by using the empirical cycle (De Groot, 1961), and scaffolding (Granott et al., 2002; Van de Pol et al., 2010; Van Geert & Steenbeek, 2005). Teachers learned to use open and student centered (Oliveira, 2010) questions during scaffolding, and learned how these questions, in combination with the use of the empirical cycle, can help enhance pupils' scientific reasoning skills. To help incorporate the three strategies, the teachers received flashcards with the empirical cycle described combined with examples of good questioning (appendix A). Secondly, during the coaching session following a lesson, the teachers had the opportunity to put this knowledge into practice and to reflect on their own behavior, with the focus on the use of the three strategies and on the effect that the strategies had on the teacher's behavior, children's thinking and reasoning and the interaction between teacher and pupils. The coach selected four or five fragments, typical for a particular teacher's behavior and interaction with the pupils during the lesson, from the video to be discussed by the teacher and coach. They discussed the video fragments of the lessons, with a focus on the teacher's interaction with the children and the structure, assistance and opportunities the children were provided with during the lesson. Special attention was paid to successful remarks which elicited better interaction with pupils, and which showed the effect the interaction had on pupils' reasoning skills. Teachers were encouraged to show more of this interaction in their next lessons. Thirdly, teachers' intrinsic motivation was stimulated by letting them formulate and work on their own learning goals before the coaching sessions.

5.2.4 Variables

The video-recordings of the lessons (pre-intervention, post-intervention and coaching lessons) were used to objectively determine the effect of the coaching on the actual questioning of the teacher (the variable: 'teacher's scientific reaso-

ning eliciting questioning', from now on TSEQ) and on pupils' scientific reasoning (the variable: 'pupils' level of scientific reasoning, from now on PLS'). With this aim, coding schemes were developed for coding teachers' and pupils' utterances. The teachers' coding scheme was developed using literature that stressed the importance of the empirical cycle and of questioning for science learning (De Groot, 1961; Dejonckheere et al., 2009; Engel & Randall, 2008; Oliveira, 2010), and focused on teachers' scientific reasoning eliciting questioning. The videos were coded in the following order: first, the main verbal statements were coded to distinguish all questions. Secondly, all questions were, as much as possible, classified in terms of the steps of the empirical cycle (table 6) they represent. Follow up questions and other questions were coded in a separate category, not as a category of the empirical cycle. For this study, the sum of all forms of scientific reasoning eliciting questions was used as the main variable. The only exception was that the knowledge questions (code 1) were left out of the calculation because these are not questions that stimulate children's thinking.

Table 6 | *Coding scheme teacher's scientific reasoning eliciting questions (TSEQ)*

code	Question type	Example
1	Knowledge questions	From what material is a pencil made?
2	Prediction questions	What do you think will happen if you throw a pencil into the water?
3	Research design questions	What materials do we need for this experiment?
4	Observational questions	What do you see? What happens?
5	Questions about the explanation	How do you think this is possible?
6	Follow up questions	What do you mean? Can you explain the answer?
7	Other questions	Can you all sit down?

The pupils' coding scheme with regard to their level of reasoning was developed using Skill theory (Fischer & Bidell, 2006), with the aim of analysing children's scientific reasoning skills, i.e., the level of verbal performance of children in

scientific tasks. Skill theory describes the developmental cycles of levels and tiers of skills during the human lifespan. In this study, only the levels that children from preschool and grade 1-2 actually use are shown (Fischer & Bidell, 2006). The coding scheme has already been used in a slightly modified way in recent research (Meindertsma et al., 2012; Rappolt-Schlichtmann et al., 2007; Van der Steen, 2014), which showed that the coding scheme could be scored reliably at sentence level. Table 7 shows the levels used, with accompanying examples of pupils' verbal utterances (answers), on that level.

Table 7 | *Coding scheme Pupils' level of scientific reasoning (PLS)*

code	Complexity level	Description	Example
0		No answer is elicited by the teacher	
1	Sensorimotor mapping	A child observes characteristics of an object	This is a blue pencil
2	Sensorimotor systems	Child states a relationship between action and result	The pencil floats because you placed it in the water
3	Single representation	Child refers to one part of the explaining mechanism	The pencil floats because it is small
4	Representational mapping	Child refers to two or more parts of the explaining mechanism	The pencil floats because it is small and light
5	Representational system	Child refers to all explaining mechanisms	The pencil floats because it is light for its size in the water

For this study, the highest level of the pupils' variable PLS was determined for every minute of the lesson. The mean PLS of each lesson was used as the main variable.

5.2.5 Data collection and data analyses

A mixed methods design was used in which both quantitative and qualitative data were gathered (Creswell & Plano Clark, 2011). Both kinds of data were systematically collected through video recordings of the classroom activities, classroom observations, and video recordings of the coaching sessions with regard to the classroom activities. The video recordings were captured with a digital camcorder, focusing mainly on the teacher to record the teacher's behavior.

Because of the labor intensive nature of coding and analysing the data, only the

first 15 minutes of the lessons were coded regarding the teachers' variable TSEQ. For the analyses of the children's utterances, the highest level of the pupils' variable PLS was coded for every minute of the lesson. The video coding program, The Observer 10.5 ("The Observer XT," 2011), was used.

The reliability of both coding schemes was assessed by training a second observer to recode 15% of the coded lessons, and to then compare this with the result of the trained observer. An inter observer agreement of 86% was found for the variable 'TSEQ' (Kappa =.85), and 86% (Kappa= .84) for the variable 'PLS', which means that the coding systems can be used reliably.

All differences between pre-intervention, intervention and post-intervention lessons were analysed. That is, the mean of the pre-intervention was compared with the mean of the intervention sessions and with the mean of the post-intervention sessions. The statistical method used is the non-parametric Monte Carlo permutation analyses (Todman & Dugard, 2001) because this test is appropriate for small samples. This test is based on the null hypothesis that no statistical difference exists between pre-intervention, intervention and post-intervention lessons of each teacher. This null hypothesis is tested by randomly permuting the eight lessons per teacher over the eight moments in time. For each random permutation, we calculate the difference between the mean of the pre-intervention and post-intervention lessons as well as the difference between the mean of pre-intervention and intervention lessons. The random permutation has been carried out 1000 times and the distribution of these results was compared with the original results. The number of times that the random permutation produces a difference that is as big or bigger than the observed difference is counted and divided 1000. The result of this division is an estimation of the exact p-value for this small sample.

Furthermore, Cohen's d is used to calculate the effect size of the intervention (Cohen, 1988). To gain more insight into the individual trajectories with its specific mutations, the slope of the amount of scientific reasoning eliciting questions is calculated, as well as the slope of the pupils' level of scientific reasoning, in order to compare the increase in empirical cycle questions and the pupils' level of scientific reasoning of each teacher with the performance of the other participating teachers. This time the serial performance variable is examined by using the slope, a linear trend parameter that describes the direction (increase or decrease) and steepness (the strength of the decrease or increase) of the changes in the variables. A level of significance of .05 is used.

To respond to the third hypothesis (with regard to the patterns of change that can be recognized between individual teacher's trajectories), the raw time series

of both teachers' amount of scientific reasoning eliciting questions and pupils' level of scientific reasoning were first smoothed, using a Loess smoothing technique, and subsequently normalized, using a linear transformation, so that both teachers' and pupils' measures were put on the same scale, both ranging from 0 to 1 (Jacoby, 2000). Smoothing the raw time series reduces the local variability, e.g. during each measurement point, while still retaining the eventual nonlinearity of the data, i.e., retaining patterns of up- or downward movements across two or more observations. Essentially, a simple regression model is also a form of smoothing, but its disadvantage is that it reduces all variability to a single, overarching trend. The Loess smoothing applied in this article reduces the short-term (micro) variability, but conserves the mid-level (meso) variability.

5.3 Results

5.3.1 Teacher's scientific reasoning eliciting questions (TSEQ)

Figure 7 | *Teacher's scientific reasoning eliciting questions (TSEQ)*
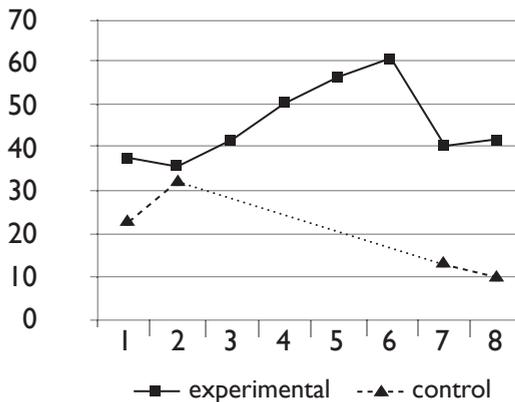
During the coaching sessions, teachers from the experimental condition ask more scientific reasoning eliciting questions than during the pre-intervention lessons, as can be seen in figure 7. This figure shows that the amount of TSEQ for the experimental group increased significantly, from 36.5 for the pre-intervention sessions to 52.3, the mean of the four intervention sessions ($p=.02$), with a medium effect size ($d=0.77$). This effect is no longer observable two months later, during the post-intervention lessons. In comparison with the pre-intervention lessons, there is no significant increase observable during the post-intervention lessons ($p=.32$).

The starting level of TSEQ of the control group is not significantly different from the starting level of the experimental group ($p=.16$). However, the teachers from

the control group show a significant drop in TSEQ from pre-intervention to post-intervention (p< .001), as can be seen in figure 7. During the pre-intervention lessons the teachers, on average, ask 28.5 questions during the first 15 minutes of a lesson, which decreases by 16.5 leaving 12 questions during the post-intervention lessons. When comparing the results of the experimental group with the control group of the pre- and post-intervention conditions, the intervention group remains on the same level, whereas the control group shows a significant drop of 16.5 on average. This difference between the two groups from pre- to post-intervention in the level of TSEQ is significant (p=.03). Although the experimental group did not show improvement during the post-intervention, they clearly did not diminish their amount of teacher's scientific reasoning eliciting questions either, as did the control group.

Because not all lessons have the same topic, and because some lessons are more suitable than others for provoking questions, the proportion of the empirical cycle questions to the over all questions was also calculated. This proportion however showed the same pattern as in figure 7, where the absolute amount of empirical cycle questions is shown: a significant increase from pre-intervention to intervention lessons (p=.01), no significant increase from pre-intervention to intervention lessons (p=.30), and a significant difference between the experimental and the control group (p=.01).

Hypothesis 1a can be partly confirmed by these results. The intervention has had a significant effect on the amount of questions asked, but the effect does not seem to last beyond the intervention. However, hypothesis 1b can be confirmed. In comparison with the control group, the experimental group's number of questions stays at the pre-intervention level, whereas the amount of questions of the control group drops to a lower amount. The fact that the experimental group did not drop below the pre-intervention amount gives an indication for a longer lasting effect of the intervention (hypothesis 3).

5.3.2 Pupils' level of scientific reasoning (PLS)

As shown in figure 8, the mean level of pupils' PLS in the experimental condition significantly increased from 1.3 for the pre-intervention lessons to 1.8, for the two post-intervention lessons with a large effect size (p=.02; d=0.83). This effect was also visible during the intervention lessons. The difference from pre- to post-intervention between the mean of the pre-intervention (1.3) and the mean of the intervention sessions (1.8) is significant with a large effect size (p=.002; d=0.89). Regarding the control group, a significant decrease of 0.4 occurred, from average 1.3 to average 0.9 (p=.05; d=1.43).

Figure 8 | *Pupils' level of scientific reasoning (PLS)*
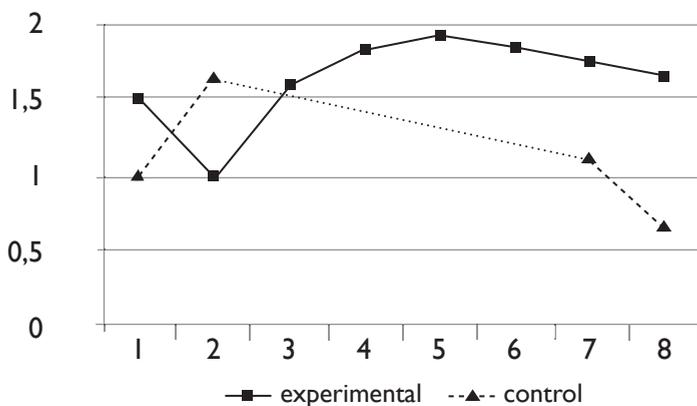
When comparing the experimental group with the control group, the experimental group shows an increase in level of reasoning of .5, whereas the control group shows a decrease of .4. The difference between the two groups is significant as well (p=.01). These results confirm hypothesis 2a and 2b, namely, that the PLS of the experimental group increases, whereas the control group's PLS does not, as well as hypothesis 3, that addresses the longer lasting effect. The pupils show an increase in their level of scientific reasoning during the program, and the experimental group shows an increase, whereas the control group shows a decrease.

5.3.3 Individual teachers' trajectories

5.3.3.1 *Slopes for the change in scientific reasoning eliciting questions for each teacher*

Table 8 | *Slopes for the change in scientific reasoning eliciting questions for teachers A-F*

	Teacher		Teacher		Teacher		Teacher		Teacher		Teacher	
Experimental group	A	p	B	p	C	p	D	p	E	p	F	p
Measurement 1-6	9,13	.03	6,52	.02	7,78	.02	6,04	.05	-0,29	.53	2,31	.38
Measurement 1-8	1,00	.39	0,65	.40	3,53	.06	4,78	.01	-2,17	.73	-0,05	.51

Slopes are calculated for all six teachers (table 8) with regard to their use of TSEQ. First, the slopes over the pre-intervention and the intervention lessons (measurement points 1 to 6), and secondly, over all lessons including the post-intervention lessons (measurement points 1 to 8) were determined. With regard

to the slopes of measurement point 1-6, four of the six teachers show a significant positive slope, i.e., they benefitted from the intervention (teacher A, B, C and D, with respectively $p=.03$, $p=.02$, $p=.02$ and $p=.05$). The change in slope of both other teachers (E and F) is not significant i.e., their results cannot be distinguished from a random pattern. The slopes of measurement points 1 to 8 show that only the positive slope of one teacher (D) is significant ($p=.01$), i.e., a significant increase in the number of questions exists. The positive slope of one other teacher (C) approaches the significance level ($p=.06$), which suggests that her slope shows a similar trend. Four teachers show no significant change.

With these results, hypothesis 4a can be confirmed. Inter-individual variability occurs, in that individual teachers will show different patterns of changes.

In summary, three patterns of change can be recognized; the first pattern entails that no behavioral change is visible during the trajectory. This pattern can be seen in the results of two teachers (E and F). The second pattern is that teachers show behavioral change until the last intervention lessons, and then fall back on their initial level, which is the case for three teachers (A, B and C). The third pattern entails behavioral change which endures over the post-intervention lessons (one teacher, D).

5.3.3.2 *Slopes for the change in pupils' level of reasoning for each teacher*

Table 9 | Slopes for the change in pupils' level of reasoning

	Teacher		Teacher		Teacher		Teacher		Teacher		Teacher	
Experimental group	A	p	B	p	C	p	D	p	E	p	F	p
Measurement 1-6	0,15	.03	0,42	.02	0,30	.02	0,03	.44	-0,14	.78	0,15	.09
Measurement 1-8	0,05	.19	0,25	.03	0,07	.27	0,11	.17	-0,07	.72	0,09	.10

Table 9 gives an overview of the slopes of the results of the participating teachers' pupils' PLS, again calculated separately during the pre-intervention and the intervention lessons (measurement points 1 to 6), and during all lessons including the post-intervention lessons (measurement points 1 to 8). With regard to measurement points 1 to 6, the group of pupils of three of the six teachers (teacher A, B, and C, $p=.03$, $p=.02$, and $p=.02$ resp.) show a significant positive slope, i.e., they benefited from the intervention. The positive slope of the pupils of teacher F approached the significance level ($p=.09$), which suggests a comparable trend in the change of PLS towards significance. The slope of the other groups of pupils (teachers D and E) is not significant, i.e., the results of these two groups of

pupils are not distinguishable from a 0 slope, indicating random fluctuation. A look at the slopes of measurement points 1 to 8 shows that the slope of only one group of pupils (teacher B, $p=.03$) shows a significant positive slope, i.e., a significant increase in their PLS, whereas one other group of pupils (teacher F, $p=.10$) shows a trend towards significance. Four groups of pupils show no significant change over this period.

5.3.3.3 Coherence in the patterns of change in individual teachers' use of TSEQ and their pupils' PLS.

Examining the coherence in patterns of change is only meaningful if any coherence exists between individual teachers' use of TSEQ and their pupils' PLS. Therefore, correlations were calculated between the amount of TSEQ and the mean PLS for all teachers and all lessons. The overall correlation was 0.62 and the correlations for the individual teachers varied from 0.54 to 0.83, representing an intermediate coherence between teachers' and pupils' behavior. These correlations are statistically significant ($p<.001$).

Next, in order to examine the level of coherence between the patterns of change in individual teachers' use of TSEQ and their pupils' PLS, a visual inspection of the graphs was carried out. In figure 9, graph A depicts the raw data of teacher A's TSEQ and her pupils' PLS, graph B the raw data of teacher B, etc.

Coherence can be defined in two ways. First, as similarity in the direction of change, as measured between two measurement points, i.e., both variables change in a positive direction or in a negative direction simultaneously, i.e., the direction of change is 'plus plus' or 'minus minus'. 'Plus minus' or 'minus plus' are incoherent changes. TSEQ and PLS can be coherent, in the sense that they move or do not move in the same direction. Figure 9a shows that some coherence can be seen in all teacher/pupils combinations during the whole trajectory, although not in all cases (for instance combination E). Overall, the pupils in the classrooms seem to show more or less the same pattern as their teacher. In one combination, all seven changes in direction that occur are coherent (combination D), in one combination six changes out of seven are coherent (C). Three combinations (A, E and F) show five out of seven coherent changes, whereas one combination shows three out of seven coherent changes (B).

Secondly, coherence can be seen in the patterns teachers and their pupils show. Here, teacher D's pupils show an increase from pre-intervention 1 on, and teacher E's pupils show the same erratic pattern as their teacher. Teacher B's pupils follow her pattern of increase or decrease until coaching lesson 4; during the post-intervention lessons the pupils stay at a high level of reasoning, whereas
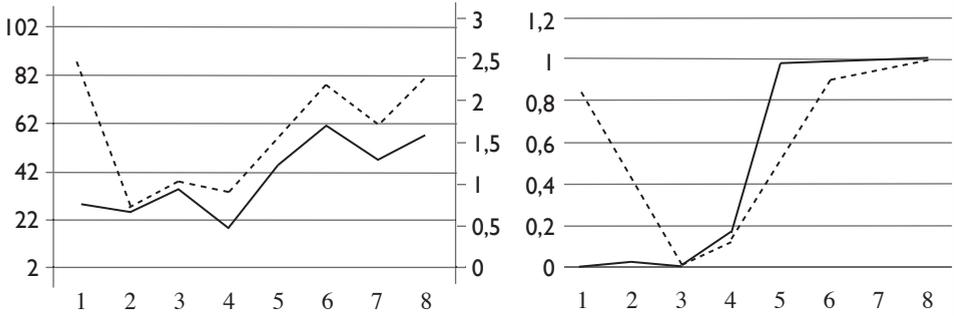
9a   9b

Teacher A

Teacher B
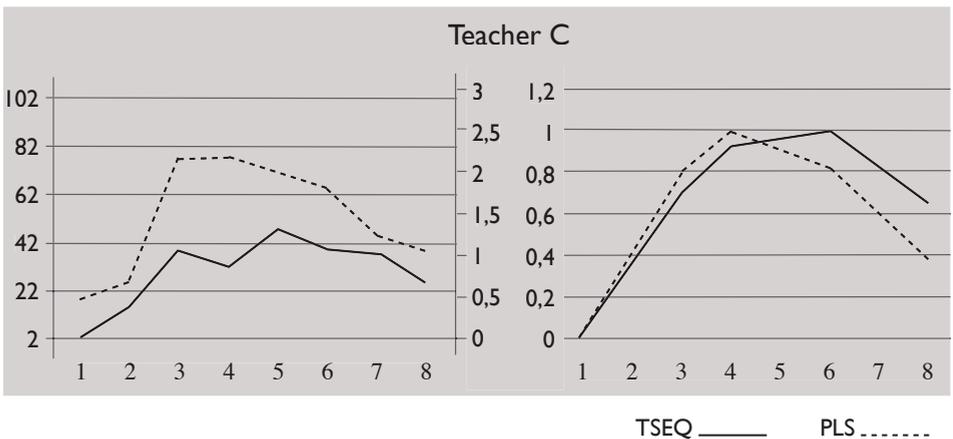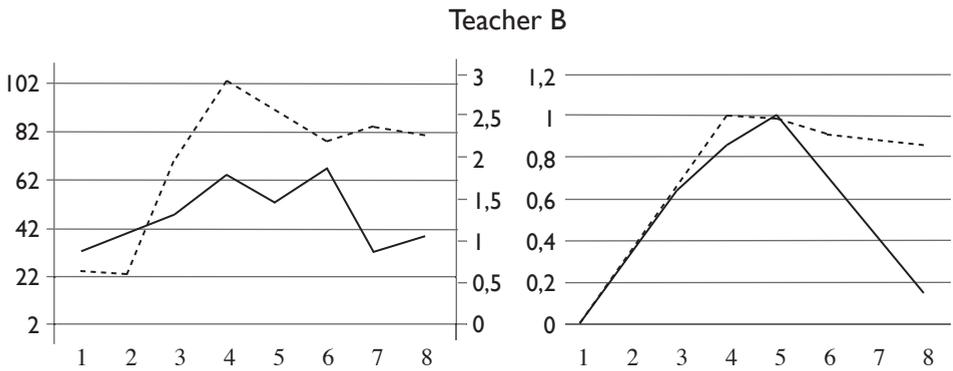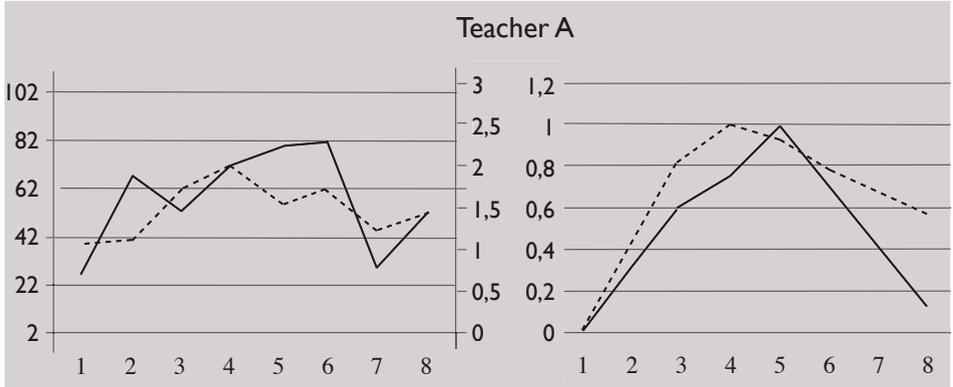
Teacher C
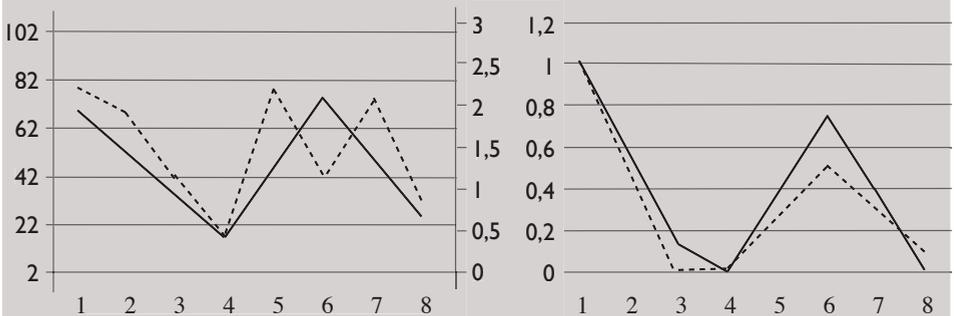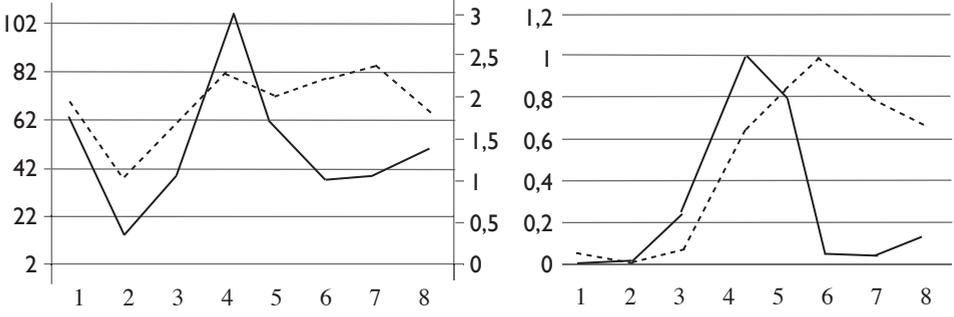
TSEQ ———   PLS ·······

Teacher D

Teacher E

Teacher F

TSEQ ——— PLS ----

teacher B herself declines. Teacher C's pupils follow exactly the same pattern as their teacher: increase during the coaching lessons, decrease during the post-intervention lessons. Teachers A and F and their pupils do not show a clear pattern.

Each way of measuring coherence gives different results. The first way, with regard to coherent changes, shows that teacher D has the most coherent results, whereas the second way, with regard to coherent visible patterns, shows that both teachers C and D have the best coherent pattern. However, both definitions lead to the same conclusion, namely, that inter-individual differences can be observed: the level of coherence differs for the different teacher-pupils combinations.

In order to determine the statistical significance, in addition to the former two coherence measures, the coherence of all teachers together with regard to their TSEQ and their pupils' PLS was calculated by means of a correlation of the slopes for both measurement points 1 to 6 and measurement points 1 to 8. Despite the individual differences, the slopes are coherent with correlations of respectively .64, representing an intermediate coherence and .37, representing a weak coherence. These correlations show that the overall trends in the teachers' questions and the overall trends in the pupils' answers or reasoning are related to one another. This correlation is not statistically significant with $p=.27$ for the correlation of the slopes for measurement points 1 to 8; however, it approached significance for the correlation of the slopes for measurement points 1 to 6 with $p=.07$.

To further investigate the coherence in the interaction patterns between the individual teachers and their group of pupils, we reduced the short-term (micro) variability by means of using smoothed, normalized graphs (figure 9b, teachers A-F). A slightly different picture can be observed in comparison with the raw data in figure 9a. The following patterns can be distinguished: two teacher/pupils combinations (C and E) show the same pattern during the whole trajectory. One combination (D) shows the same pattern only after both pre-intervention lessons. The graphs of combinations A and B show that the levels of teacher and pupils stay at approximately the same level during the first six measurement points of the coaching trajectory. Their pupils, however, stay at a higher PLS during measurement point seven and eight, at the same moment that the teachers' TSEQ drops. Combination F shows yet another pattern: the teacher's amount of TSEQ increases only marginally during the trajectory, and decreases again at the end, whereas her pupils' level of reasoning increases. In these latter three cases (A, B, and F) the pupils' PLS stays at a higher level, even if their teachers' TSEQ decreases.

Comparing the patterns found in these smoothed graphs with the averaged

results for the separate variables TSEQ and PLS, as shown in figure 7 and 8, it can be seen that the results of the combinations A, B and F are in line with the averaged results of the separate variables. Combination C shows a different pattern: the TSEQ and the PLS show the same pattern during all measurement points 1-8. In these four combinations (A, B, C and F) the pupils' PLS at the end of the trajectory is clearly above the starting level. Combination E shows another pattern: both teacher's TSEQ and pupils' PLS are erratic, and are just slightly in line with each other. Combination D shows a unique pattern as well: the teacher starts at a low level of TSEQ, whereas her pupils' PLS starts at a high level. During the trajectory, the TSEQ and the PLS gradually merge.

Hypothesis 4b cannot be confirmed with these results. We expected an increase over time in both TSEQ and PLS in individual teachers. We can conclude that a form of overall coherence exists with regard to the patterns of change observable in individual teachers and their accompanying group of pupils according to the raw data. The smoothed data show this same coherence, but in addition, they give more information about the patterns that can be found: they show that an increase over time in both variables can only be seen in combination C. Three combinations (A, B and F) show an increase in PLS only; the other combinations (D and E) show no increase at all.

5.4 Conclusion and Discussion

5.4.1 Conclusion

In this study, we investigated the effects of a coaching program (VFC-T) on teachers' scientific reasoning eliciting questions and on their pupils' level of scientific reasoning. We expected teachers to increase their number of questions that stimulate pupils' scientific reasoning results (hypothesis 1a). This expectation was confirmed, as the amount of TSEQ asked by the experimental group increased significantly; from 36.5 for the pre-intervention sessions to 52.3, the mean of the four intervention sessions (p=.018), with a medium effect size (d=0.77). We also expected this increase to have a longer lasting effect, as shown in the post-measurement two months after the program (hypothesis 3). This expectation was not confirmed. During the post-intervention lessons, the TSEQ of the experimental group diminished to about the same amount of questions as were asked during the pre-test lessons.

However, the hypothesis that the results of the experimental group would differ from the results of the control group (hypothesis 1b) was, in a sense, indirectly confirmed: a significant drop in TSEQ could be seen from pre-intervention to post-intervention lessons for the control group. This finding is consistent with

declining developmental trends in quality indicators of teaching practices during a school year as a relatively ubiquitous phenomenon, which has been reported in earlier studies using observational measures of teachers behavior (Evertson & Veldman, 1981; Opdenakker, Maulana, & den Brok, 2012; Stroet, 2014). A possible explanation for the phenomenon that the control, as well as the experimental group, showed a decline in comparison to the earlier levels of asking questions could be that the intervention was carried out in March and April, which implicates that the post-intervention lessons were recorded in June, at the end of the school year. During the end of the school year, teachers tend to give more emphasis on social activities, which implies that teachers and students are less focused on high level teaching and high-level intellectual performance than during the rest of the school year. The fact that the level of questions in the experimental group was considerably less affected by this end-of-year phenomenon than the control group could elucidate the results of hypothesis 3, and provides a justification for the long term effectiveness of the intervention. This effectiveness is found in the *difference* between the final levels of reasoning eliciting questions between the experimental and the control group.

Another explanation for the differences between the experimental and the control group could be that an observer effect is present; i.e., the effect that individuals improve or modify an aspect of their behavior in response to their awareness of being observed. In order to partially reduce this effect, we did not communicate to any of the participants which variables would be used to measure the effectiveness of the intervention.

Secondly, we expected pupils in the experimental group to show an increase in the PLS during the program, as well as afterwards, demonstrated by the post-measurement two months after the program (hypothesis 2). This expectation could be confirmed. In addition, it is very clear that the pupils were, in fact, much less influenced by observer effects or end-of-year effects than their teachers, since they do not show a decline towards the end of the year.

Thirdly, the expectation was that the patterns of change in the relationship between teachers and pupils would show considerable inter individual variability (hypothesis 4a). This expectation was clearly confirmed. Three patterns can be recognized, namely, no behavioral change at all, behavioral change until the last intervention lessons, and behavioral change which endures over the post-intervention lessons (Wetzels, Steenbeek, & Van Geert, 2015). Finally, we expected coherence with regard to the patterns of change observable in individual teachers and their accompanying group of pupils (hypothesis 4b). The expectation with regard to the coherence in the patterns of change could be confirmed, first by

means of the overall correlation and the correlations for the individual teachers between teachers' use of TSEQ and their pupils' PLS, secondly by means of correlations of the slopes of the teacher-and pupils-variables, and finally, by means of additional coherence measures. As predicted in hypothesis 4a, the patterns of coherence also showed clear inter-individual variability. The dominant pattern that three of the six combinations show, however, is that pupils have learned to reason at a higher level at the end of the coaching trajectory, and have benefited from the way teachers elicit scientific reasoning in the earlier coaching sessions, even though the teachers regress during the final sessions. That is, the intervention seems more sustainable for pupils than for teachers, although the sustainability for teachers depends on whether an end-of-the year or observer effect exists or not.

5.4.2 Limitations and future research

This multiple case study was conducted with a small group of teachers originating from two schools and one coach. The small number of teachers is a direct consequence of the fact that we wanted to answer our research question by means of a labor-intensive, but ecologically valid method, namely, collecting observation data in an authentic situation, and by means of a time serial study of repeated, observed lessons. Because of this methodological choice, the number of participants is, by definition, small. It is clear that the number of observed classes should be increased during further, comparable in-depth studies of authentic teaching situations and reasoning of children.

Furthermore, large inter-individual variability between teachers, as well as intra-individual variability within teachers, were observable in both variables. These results resemble findings in research that focuses on linguistic variables in teacher-pupils interaction (Menninga, van Dijk, Wetzels, Steenbeek, & Van Geert, 2015). Tentative patterns hopefully can be confirmed by future research that focuses on larger scale data, with an emphasis on overall results, as well as individual results, both quantitative and qualitative (Creswell & Plano Clark, 2011). This combination can give more insights regarding the elements that affect the effectiveness of an intervention. For instance, in this study, one teacher/student combination (F) hardly benefited from the intervention. The fact that an intervention is effective at a group level does not imply that it is effective for each individual case in the group. More insight is needed regarding the elements that are responsible for deviant individual results (Barlow & Nock, 2009; Molenaar & Campbell, 2009; Rose et al., 2013).

An important limitation of the current study is that teachers were only coached for four weeks, without any follow-up coaching. It is questionable whether this

relatively short coaching period was long enough in order to yield sustainable results; however, long interventions are expensive and sensitive to a lot of attrition. It should be mentioned that other studies (Fabiano et al., 2013; Garet et al., 2001; Han & Weiss, 2005) do suggest that longer lasting trajectories are more effective in this respect. Longitudinal research on the effect of the intervention, with perhaps one or two follow up coaching sessions each year, might reveal that follow up sessions do contribute to the sustainability of an intervention (Dishion & Stormshak, 2007; Guskey & Yoon, 2009; Sandholtz & Ringstaff, 2013).

Another limitation was the way students' performance was measured, namely, only at a group level instead of on both a group level and individually. In addition to studying the level of aggregation of the class, it is of course also very interesting to study the answers at an individual level because ultimately the effect of an intervention must ideally result in an effect in every individual pupil. Focusing on the class level conceals the fact that the intervention might be fine for a rather small subset of students, but entirely miss a number of other students, and possibly even miss the students who could most benefit from the intervention. Regrettably, focusing on individual levels was not possible because of the limited amount of answers of each pupil, and because individual answers in the group could not always be allocated to one specific pupil. The results, therefore, can only be used to measure the effectiveness of the intervention for teachers, individually and on the group level, and not on the level of individual pupils. Therefore, an observational study on the level of individual pupils is a recommendation for the future.

In this study, we limited ourselves to reporting teachers' and pupils' verbal utterances, and more specifically, to teachers' questioning and pupils' level of reasoning. These are not the only variables that can be used to observe changed behavior in the classroom. In more fundamental research, Meindertsma (2014), for instance, uses a variable that specifies the level of openness teachers employ in their utterances, and other researchers examine pupils' nonverbal behavior during reasoning tasks (Goldin-Meadow & Alibali, 2002; Hoekstra, van der Steen, Cox, & Van Geert, n.d.; Thelen & Corbetta, 2002).

We suggest addressing these aforementioned limitations in future research by using a similar real-life design and in-depth study of authentic teacher and pupil behavior, complemented with firstly, an additional number of teacher-pupils combinations, secondly, an increase in the duration of the intervention, and finally an extension of the observed variables including teacher openness and nonverbal reasoning behavior of the pupils.

As is extensively described in chapter 3, a teacher and a school are each complex dynamic systems, which implies that many aspects dynamically determine the effectiveness of an intervention. During the intervention, the coach has to adapt the intervention to the participants' circumstances and specific needs, paying close attention to the diverse ways all the aspects are intertwined and influence each other. In turn, school boards should not treat an intervention as an isolated tool, the effect of which entirely resides in the tool itself. Instead, they should treat an intervention as a tool, the effectiveness of which, is highly dependent on the way this tool is used under the circumstances and concrete contexts, as, for instance, the motivation of the participants. School boards should make efforts to prepare and maintain the local and current context of the intervention, i.e., the school, the teacher and its surroundings, in an optimal way, so that effectiveness can be optimized.