Res Sci Educ (2013) 43:551–578
DOI 10.1007/s11165-011-9276-8

The Development of Understanding of Evidence in Pre-University Biology Education in the Netherlands

Herman H. Schalk · Joop A. van der Schee · Kerst Th. Boersma

Published online: 30 December 2011

Abstract Ensuring the quality of investigations requires the understanding of procedures by which empirical evidence is obtained. This can be interpreted as becoming aware of and using criteria for evidence in one's mental structure. The question is whether this process can be observed in practice. In two schools for pre-university education where 11th grade students were working in small groups on open investigations in biology, all conversations in class -with or without the teacher participating- of eight groups (17 students) were audio-taped and transcribed. All utterances concerning the quality of investigations (3943 in total) were analyzed using five categories: problematization, description, explanation, generalization and application. Half of the students received written feedback twice, the other half paused their own investigations to carry out four specially designed reflection tasks. Talking about the reflection tasks as well as having the teacher present in conversations about investigations have shown to stimulate the spiral of description, explanation and generalization. Students who did the reflection tasks explained and generalized significantly more than students who did not. Still, the majority of the explanations and generalizations came from the teacher. Implications for the role of reflection and the role of the teacher in developing procedural understanding are discussed.

Keywords Biology education · Understanding of evidence · Procedural understanding · Quality of investigations · Reflection-in-action · Reflection-on-action

Introduction

The stance that students should do practical work in science has been advocated for a long time (see e.g. Lunetta et al. 2007). The goals pursued by letting students carry out practical

H. H. Schalk (✉) · J. A. van der Schee
CETAR Centre for Educational Training, Assessment and Research, VU University, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands
e-mail: h.h.schalk@vu.nl

K. T. Boersma
Freudenthal Institute for Science and Mathematics Education, Utrecht University, Utrecht, The Netherlands

work, however, can vary to a great extent (Abd-El-Khalick et al. 2004; Hodson 1993, 1998; Hofstein and Lunetta 2004; Séré 2002). If practical work is narrowed down to doing investigations or inquiries, a dichotomy between 'inquiry as means' and 'inquiry as ends' can be distinguished (Abd-El-Khalick et al. 2004). The first emphasizes the development of science content by doing inquiries, whereas the latter focuses on the development of understanding of the skills and processes of science. In open-ended investigations (i.e. investigations in which there is not one 'correct' answer and there are many routes to a valid solution), substantive content knowledge as well as procedural understanding appears necessary to succeed (Glaesser et al. 2009b). When students carry out open-ended investigations in biology education, they will have to draw on their substantive knowledge of the subject investigated, on their ability to use practical skills and on their procedural understanding, for instance, about how to construct a fair test or how to use evidence in reasoning (Millar et al. 1994), something that Klahr and Dunbar (1988) call 'dual-space searching'.

Having observed students working on open-ended investigations in Dutch secondary schools, in lower as well as in higher grades, we notice that they are very much committed to their work: they conceive it as something of their own, and they really want to get a good final result. Students work hard, but the quality of their work is often poor. It lacks understanding of criteria for gathering and evaluating the quality of empirical evidence, the understanding of how to proceed, the so-called 'procedural understanding' (Millar et al. 1994; Glaesser et al. 2009b). Procedural understanding can be seen as having a knowledge-base upon which scientists draw when they design and evaluate (evidence coming from) experiments, and therefore understanding plays an important role in scientific reasoning (Roberts 2001).

The recognition of this problem led us to develop teaching and learning strategies that would enhance the understanding of scientific evidence by students in the subject of biology in upper secondary and pre-university education in the Netherlands. In this paper we report about one aspect of the above mentioned strategies: the role of reflection in the development of understanding of evidence.

Before describing our research questions, methods and results, we sketch the background of our research and our theoretical framework. First, we elucidate the origin of the problem: the poor quality of students' open-ended investigations. Second, we discuss the role of procedural understanding and introduce the 'concepts of evidence' as a possible contribution to the solution. Third, we sketch a theoretical framework with which we can analyze the learning and teaching of understanding of evidence. Fourth and finally in this introductory section, we distinguish the two ways in which reflection was used in our teaching and learning strategies.

Quality of Open-Ended Investigations

In textbooks, doing investigations is primarily aimed at illustrating theory (Lunetta et al. 2007), even when it is moulded as an experiment. Recipe-like instructions keep most students away from thinking about the purpose of the experimental design or about how many times to repeat a measurement (Schauble 1996; Schauble et al. 1995). This is mirrored in the difficulty students have with the understanding of the nature of theory and experiments (Duveen et al. 1993) or with using a scientific attitude towards investigative tasks (Millar et al. 1994). Chin and Brown (2000) taped US grade 8 students during class group laboratory activities. These demonstrated a surface learning approach, where the students' explanations were mere reformulations of the initial questions. Reigosa and

Jiménez-Aleixandre (2007) observed how a stereotypical school culture obstructs genuine problem solving in the science laboratory.

However, an emphasis on inquiry skills is not the solution for the lack of procedural understanding either, because a focus on the performance of skills can exclude understandings about inquiry and the nature of science, as Lederman states in his contribution to an article about international perspectives on inquiry in science education (Abd-El-Khalick et al. 2004). Attempts have been made to enhance students' procedural understanding, with varying results (e.g. Buffler et al. 2001; Rollnick et al. 2002; Trumbull et al. 2005; Van Rens et al. 2004).

As the problem of insufficient procedural understanding appears to be widespread, it is not surprising to observe that Dutch students have difficulty in designing investigations themselves (Smits et al. 2000) or formulating research questions (Van der Schee and Rijborz 2003). In biology textbooks concepts like 'fair testing' and the relation between question and conclusion often receive some attention, but the understanding and application of the concepts of validity and reliability recieve less attention (Van Rens 2005).

An analysis of the quality of 169 student reports from six different Dutch high schools led to the conclusion that in the majority of reports only a restricted understanding of evidence was displayed (Schalk et al. 2009). Taking into account that these reports came from schools where students *did* open-ended investigations – which is surely not the case in every school – we assume that the overall situation in the Netherlands can benefit from teaching and learning strategies that enhance the understanding of evidence.

Procedural Understanding

Understanding what counts as evidence in an investigation requires the ability to reason scientifically. Although adults outperform adolescents and children in scientific reasoning (Kuhn et al. 1988), even they do not always succeed in generating and interpreting evidence, especially in situations that resemble real-world settings (Schauble 1996). Scientific reasoning comprises hypothesis generation, experiment design and evidence evaluation (Klahr 2000). However, answering a question about what counts as evidence does not only require a reasoning about whether experimental data support a certain conclusion, but also knowledge and understanding about how one creates circumstances to obtain valid and reliable evidence. For instance, to generate a hypothesis, specific knowledge of the scientific domain is needed, but also knowledge about how to formulate testable hypotheses, possibly including competing hypotheses, is needed. To design experiments and to investigate the natural world, knowledge about appropriate apparatus and investigative methods as well as knowledge about the quality of measurements and sample taking is required. Also, in order to evaluate evidence, not only logical inferences between research question, hypothesis and obtained results have to be drawn, but also the question has to be answered about how accurate the measurements and how reliable the results are. These latter aspects of all three elements of scientific reasoning can be put under the heading of 'procedural understanding'. In scientific reasoning research these aspects are often present, but mostly not considered explicitly (e.g. Lehrer et al. 2008). Gott and Duggan (1995a, 1996) state that procedural understanding, together with conceptual understanding, constitutes the foundation of 'thinking scientifically' and scientific literacy.

We consider scientific reasoning at least to some extent equivalent to critical thinking. Therefore, we can cite Bailin (2002, p. 373) with approval when she states: "It is the adherence to criteria which is at the centre of thinking critically, and giving attention to explicating and applying the relevant criteria must be at the centre of attempts to foster critical thinking in science." And in our view, such criteria can be derived from a set of concepts of evidence, which have been described in detail by Gott et al. (n.d.), to cover procedural understanding of all aspects of scientific inquiry.

In their description of what they mean by concepts of evidence, Gott and Duggan (1995a) emphasize that the concepts do not refer to the skill of, for instance, taking measurements, but to the decisions that have to be made about what measurements to take, how to take them and how many. This shows that concepts of evidence contain cognitive as well as metacognitive components. These concepts can be used to make this attitude operational: they can be used as tools to guard or ensure the quality of scientific inquiries (Roberts and Gott 2002). Ensuring the quality of scientific inquiries is precisely what open-ended research projects in Dutch biology education need.

From these concepts, we have derived a set of requirements or criteria for any inquiry. In order to apply the set of criteria to the situation of Dutch biology education, some adaptation was necessary. To be useful in Dutch biology education, they should fit all or at least most of the types of student inquiries in biology. One type of inquiry that is common in Dutch biology education, but seems to be lacking in the approach of Gott et al. (n.d.), is the descriptive study of, for instance, an ecosystem or the behaviour of animals in a zoo. Application of the criteria should therefore not be restricted to hypothesis testing investigations, but should also cover the design and evaluation of descriptive studies. Therefore, we have attempted to formulate the criteria in such a way that descriptive studies as well as hypothesis testing investigations are included. The result is a set of 23 criteria (see Table 1), not all of which will be applicable to every inquiry, but in our opinion every inquiry can be evaluated with a subset.

Developing Procedural Understanding

Recognizing that criteria for the quality of investigations exist may be the first step in the understanding of evidence, but a strategy should be added to get students to learn it. The sociocultural perspective of apprenticeship (Collins et al. 1989; Lave 1997) provides a basis for such a strategy. In science, meanings and explanations and theories are negotiated within the scientific community. In science education, this is not different (e.g. Mortimer and Scott 2003). Students will have to interact verbally with each other to develop understanding of the criteria. As previously cited, Bailin (2002) stresses that it is the adherence to criteria which is at the centre of thinking critically, and is at the centre of ensuring quality of investigations. Therefore, if students learn to ensure the quality of their investigations, we assume the following components will play a role:

- They should learn about criteria for the quality of investigations and determine how these criteria should be applied in an investigation.
- They should apply these criteria in their own investigations; after all, they should learn to do 'good' investigations and that requires carrying out investigations themselves.
- After carrying out an investigation, they should be able to evaluate whether the criteria have been applied correctly. In other words: they should be able to reflect on their investigations and judge whether it was good enough.

Consequently, ensuring the quality of investigations is a combination of knowing the criteria for investigations, applying the criteria by carrying out an investigation and reflecting on the criteria and the quality of the investigation. This is complex, and if

Table 1 Twenty-three criteria to evaluate the quality of investigations

Related to the research question

1. A research question:

- consists only of unambiguous terms and formulations;

- is sufficiently specific and confined.

2. It is important to distinguish whether the research question demands a description or the testing of a hypothesis.

Related to the hypothesis

3. A hypothesis should be formulated that fits the research question: it is an expected result or a possible explanation.

4. The hypothesis should be testable.

5. It is (nearly) always possible to postulate more than one hypothesis.

6. Based on the hypothesis (and a number of assumptions), a prediction is formulated about which observations or measurements can be expected (if the hypothesis [and the assumptions] is true, then …).

Related to the design of the enquiry

7. In descriptive studies, the features to be observed should be stated explicitly.

8. In hypothesis testing research, the dependent and the independent variables should be identified.

9. In hypothesis testing research, all other variables that may have an influence should be identified and kept constant if possible.

10. In hypothesis testing research, a 'control experiment' should be included.

Related to observations and measurements

11. Observations and measurements should not influence the outcome of those observations or measurements themselves, and any unavoidable influence must be made explicit.

12. The range and the intervals of the chosen values of the independent variable should match the expected variation of the dependent variable.

13. The sensitivity of the instrument should be appropriate to the measurements to be taken.

14. The sample should be representative of the population

- large enough

- random or stratified, and the method of sampling should have nothing to do with the subject of research. The number of measurements or observations should be large enough.

15. Aberrant or anomalous results should be examined to decide whether they should be used further (can they be qualified as outliers?).

16. If measurements are averaged, it should be in accordance with the content.

17. One should only speak of differences between measurements if the difference is statistically significant.

Related to conclusions and explanation

18. The conclusions should match the research question and (in the case of hypothesis testing) the hypothesis.

19. A correlation is not the same as a causal relationship.

20. It should be stated whether the explanation is causal (related to the causes) or functional (related to the consequences).

Related to evaluation

21. If an enquiry is not carried out validly and reliably (in other words: if the preceding conditions are not met), the results are open to dispute.

22. If possible, the results of the enquiry should be confirmed by data from other research.

23. The conclusion should be compared with accepted ideas (theories), common sense and experience.

students are to learn it, we can assume they will not learn it by doing it once. The components of this learning process should be addressed several times.

The relation between the components is depicted in Fig. 1. In this figure the components are connected with two-ended arrows, with a large and a small tip. The large tips indicate the most obvious way of proceeding through the triangle. The other way around is possible as well, as indicated by the small tips. The interaction takes place at every stage.

The scheme of Fig. 1 puts emphasis on the activities students should carry out in learning to ensure the quality of their investigations, but is not clear about the way the criteria are 'internalized', i.e. how students take in the criteria in their mental structure and develop procedural understanding. For this purpose the model of the 'spiral formation of mental actions' of Galperin can be used, as put in a spiral scheme by Arievitch and Haenen (2005; Fig. 2). Galperin worked out the concept of internalization, a core concept in Vygotsky's (1978) learning theory. Galperin (1969, 1989) argued that the internalization of an action starts at the material level of concrete objects, then it is put into words at the verbal level in forms of overt or covert speech, after which it can be raised to the mental level of images, associations and concepts.

Applying this model to doing investigations it can be read as: orienting on the investigation (what do you already know and what can you already do?), moving on to the material level of carrying out an investigation, moving further on to the verbal level of talking about the investigation to arriving at the mental level of thinking about an investigation. For instance, in an investigation about the factors effecting plant growth, students orient themselves by discussing which variables might influence plant growth. They start to vary temperatures, colours of light, amounts of water and nutrients. In a discussion, they find out that to make sense of the results they should compare situations in which only one variable varies. By generalizing this criterion – 'the dependent and independent variables should be identified' – it can be internalized. In a subsequent investigation it then becomes a part of the orientation: what factors are we going to vary and what factors will be kept constant?

The spiral form indicates a progressing internalization and control of an action, with a simultaneous progressing quality. The progress is also shown by an improved ability of the students to orient themselves, which leads to a better understanding and implementation of the action. Step by step students become aware of the 'ins and outs of an action', and the action becomes gradually abbreviated and automated. So, this model assumes that an action that is first conducted at the material level can be 'conceptualized'. In our example: manipulating with several concrete factors in an investigation can lead to the concept of a fair test, which is one of the concepts of evidence. This is an important assumption to mention, because it can give clues for the arrangement of learning situations.

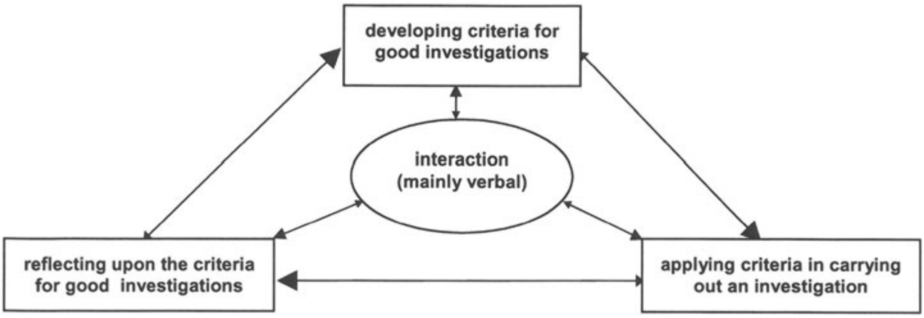
developing criteria for good investigations
interaction (mainly verbal)
reflecting upon the criteria for good investigations
applying criteria in carrying out an investigation

Fig. 1 Components of learning to ensure the quality of investigations

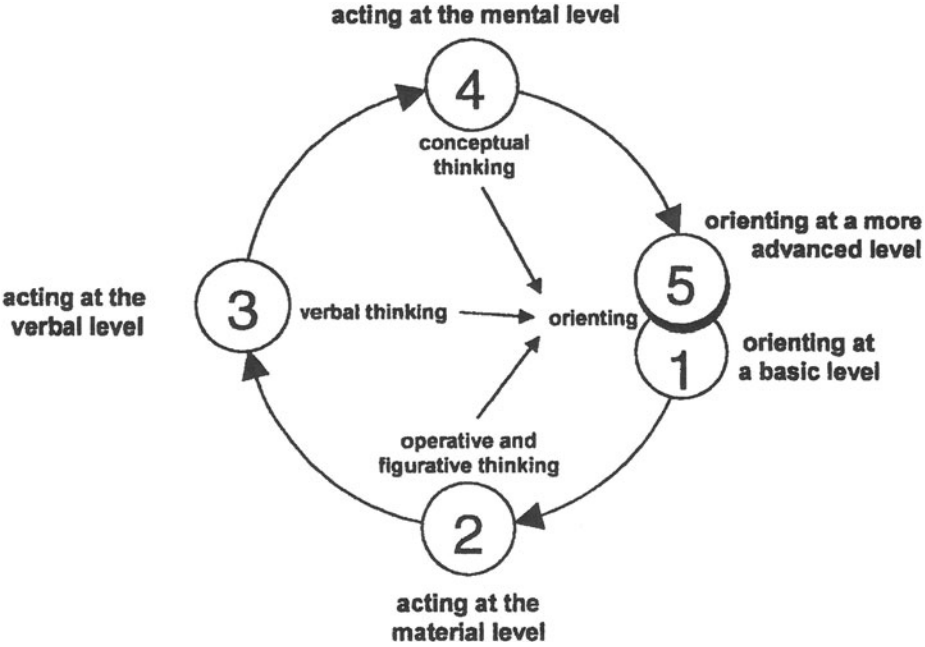
acting at the mental level
4
conceptual thinking
orienting at a more advanced level
5
acting at the verbal level
3
verbal thinking
orienting
1
orienting at a basic level
operative and figurative thinking
2
acting at the material level

Fig. 2 Galperin's 'spiral formation of mental actions' according to Arievitch and Haenen (2005)

Then, how can the processes depicted in Fig. 1 and the model of Fig. 2 be combined to understand how students internalize the criteria? The simplest way seems to be: using the criteria, talking about the criteria and internalizing the criteria. However, since verbal interaction takes place in every step in Fig. 1, and internalization cannot be observed directly, it is not that simple. We think a combination of the two processes can be found in distinguishing three levels of talking, as described by Mortimer and Scott (2000, 2003) in their analytical framework for interactions in science lessons.

- *Description* relates to utterances about specific objects, systems or phenomena in terms of their composition, value etc., i.e.: "You cannot tell whether these plants grew faster because they got more light or because they got more water."
- *Explanation* relates to utterances that bring in some theoretical notion or model to explain a specific phenomenon, i.e.: "Only when just the amount of light is different you can see its influence."
- *Generalization* relates to utterances which are not related to one specific context, i.e.: "You should vary only one variable at a time."

Figure 3 shows the combination of Mortimer and Scott's framework with the model of Galperin, where the initial orientation is called *problematization* and the subsequent orientation is called *application,* because at that point concepts are applied to a new situation. In this study we use this combination as a basis for analyzing student interaction.

We use this combined model as part of our framework to understand the ways in which students internalize the criteria. It might suggest a simple or linear - 'idealized' - way of progressing from the material through the verbal to the mental level, but there is no assumption in the model that the transitions between the levels will always occur or will occur automatically. If that were the case, why bother about an adequate teaching and

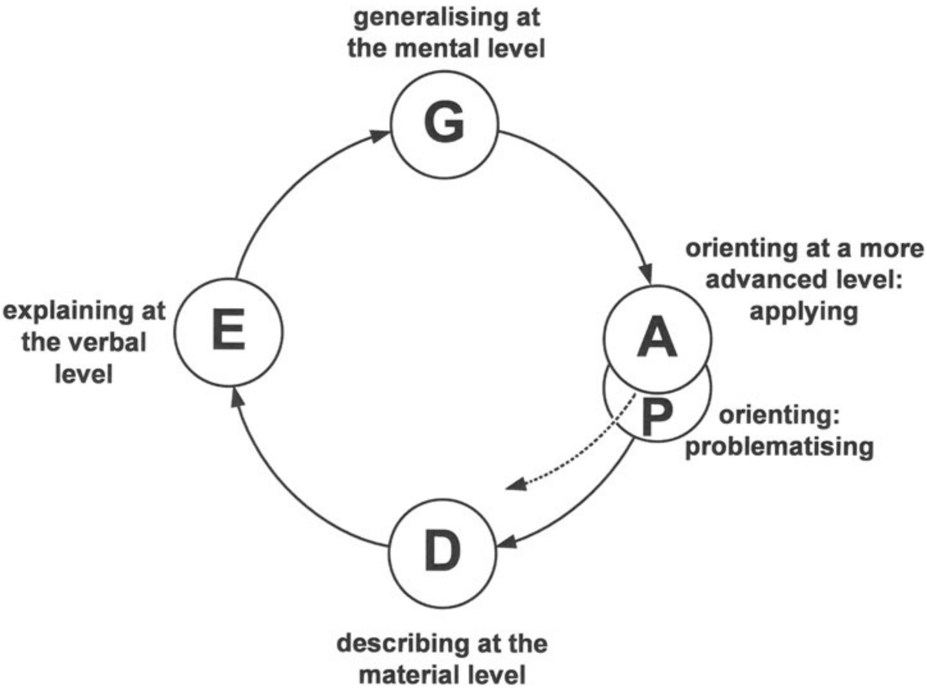
generalising at the mental level
G
orienting at a more advanced level: applying
A
P
orienting: problematising
explaining at the verbal level
E
D
describing at the material level

Fig. 3 Galperin's 'spiral formation of mental actions' applied to the development of understanding of evidence

learning strategy? Students will encounter trouble explaining a phenomenon or generalizing an explanation, as any teacher will confirm. Additionally, it is not an individual process. The interaction central in Fig. 1 has been left out in Fig. 3, but that does not mean that the internalization takes place in isolation. And although we are aware that it is a model that tries to reflect reality and it is not reality itself (Giere 2001), our main goal is not to evaluate this model, but to understand and optimize a part of science education. We think this model can help us understand how students learn and hence how this learning might be stimulated.

Two Ways of Reflection

In the APU problem-solving model, which was used to assess the quality of pupils investigative work at ages 13 and 15 (Gott and Murphy 1987), the evaluation of method and results is an essential final step which can lead to change in research question, design or technique. In writing about this model, Gott and Duggan (1995a, p. 43) state: "This continuous reflection on the design and implementation in the light of the problem as set and the requirement of the data to answer it is the single most important factor missing in pupils' work, in all investigations and at all ages." Therefore, while doing investigations it is important that students verbalize and generalize the quality of investigations: they should *reflect* on what they are doing, that is, *reflect* on the quality of their own investigations. However, in designing *reflective* learning activities we encountered some dilemmas. These dilemmas are about the content, the initiative and the moment of reflection.

Content Should the teacher ask students to reflect on *every* selected criterion, or *only those* that are relevant to their own investigations? Both Millar et al. (1994) and Séré (2002) regard the procedural understanding as a knowledge base which (therefore) should be taught explicitly. However, for instance, is raising the concept of dependent and independent variables sensible and fruitful if a student is not handling that concept in his own investigation?

Initiative Should the teacher wait to reflect on the quality of their investigations until students question it themselves or should he/she raise questions about quality when *he/she* thinks it is relevant? The first option carries the risk that they will not reflect on anything at all, the second seems to conflict with the idea of self-guidance of the students.

Moment Should the teacher interrupt the investigations of students for a moment of reflection or should he/she search for moments in the margin of their investigations?

Within the group of teachers and researchers involved in designing the teaching and learning strategy, consensus was found about the second dilemma only: the initiative should not rest exclusively with the students; the teacher should take the initiative to reflect as well.

With respect to the other two dilemmas we could not decide whether the students should reflect on the quality of their investigations while carrying out the investigations, or should they be asked to take a step aside; that is, to reflect on the quality of another investigation, to derive criteria from this reflection, and to apply those criteria to their own investigations. These two ways of reflecting are called *reflection-in-action* and *reflection-on-action* by Schön (1983). Reflection-in-action is a process that takes place during an activity and helps to direct the activity:

> “There is some puzzling, or troubling, or interesting phenomenon with which the individual is trying to deal. As he tries to make sense of it, he also reflects on the understandings which have been implicit in his action, understandings which he surfaces, criticizes, restructures, and embodies in further action. It is this entire process of reflection-in-action which is central to the ‘art’ by which practitioners sometimes deal with situations of uncertainty, instability, uniqueness, and value conflict.” (Schön 1983, p. 50)

In short, reflection-in-action is posing questions like ‘what am I doing?’, ‘why am I doing it?’, and ‘will this lead to the desired result?’ The fact that they (principally, see Roth 2009) cannot know the answer to this last question with certainty does not prevent them from posing it. If an action is fully internalized and abbreviated to a more or less automated operation, reflection-in-action does not take place. However, students in our study are not ‘automatically’ using criteria while carrying out investigations. But even then, this type of reflection can be prompted by a ‘breakdown’ (Koschmann et al. 1998) or by a question from someone else, e.g. a teacher. So, in our study, reflection-in-action is initiated and directed by (‘breakdowns’ in) the investigations of the students or by questions from the teacher during the investigative process. Typically, reflection-on-action takes place after the activity has been finished or interrupted and poses questions like ‘what did I do?’, ‘was it good enough?’, and ‘should I do it in another way the next time?’ Or, in the case the action was carried out by others, the questions would be ‘what did *they* do?’, etc. Consequently, in our study, reflection-on-action is initiated and directed by assignments in which (parts of) other investigations are questioned.

We have studied the effects of both types of reflection by designing two slightly different strategies, the *reflection-in-action (RIA)* strategy and the *reflection-on-action (ROA)* strategy. These strategies will be described in the design section of this article.

Research Context and Focus

Our main concern is the design of teaching and learning strategies in order to enhance students' abilities to evaluate the quality of their investigations and develop procedural understanding. We consider our research as design research (Brown 1992; Collins et al. 2004; Van den Akker et al. 2006). Design research can be characterized as interventionist, iterative, process oriented, utility oriented, and theory oriented (Van den Akker et al. 2006).

The study reported here is part of a larger study about the feasibility and effectiveness of the developed strategies to enhance students' abilities to evaluate the quality of investigations (Schalk 2006). Elsewhere we have reported on other aspects of this study, particularly on the learning outcomes (Schalk et al. 2009), which we will summarize here. Before and after the project, a test, focused on the use of criteria for good investigations in the evaluation of other investigations, was administered and the articles as well as the drafts versions of the articles were analyzed with regard to the use of these criteria. The results showed (1) that the post-test scores were significantly higher than the pre-test scores, with greater differences in the ROA strategy, and (2) that the students who had had the most time in class to spend on their own investigations and had received written feedback from their teachers (the RIA strategy) were better in applying the criteria to their own investigations. Since the focus of the test was on the use of criteria in the evaluation of other investigations, we concluded that both strategies enhance this ability, but that the ROA strategy seems slightly better in this respect. To find out how both strategies contribute to the enhancement of this ability we try to analyze the ways of development of procedural understanding, and in particular the effect of the reflection tasks. Therefore, focus of this article is on the process of reflection as part of the intended strategies.

Research Questions

The main question to be answered in this article is: How do pre-university students develop procedural understanding, i.e. how do they acquire criteria for the evaluation of the quality of investigations when carrying out investigations in biology; can the spiral of description, explanation and generalization be observed?

A subsequent question is: How is the process stimulated best: by *reflection-in-action* or by *reflection-on-action*?

Design of the Study

Design research "uses ongoing in situ monitoring of the failure or success of (alternative versions of) some designed artefact (software, curricular intervention, tutoring sessions, etc.) to provide immediate (and accumulating) feedback on its viability, its 'learning theory' or 'hypothetical learning trajectory'." (Kelly 2006, p. 107). Collins et al. (2004, p. 20–21) list several aspects in which design research methodology differs from laboratory studies. An important goal is to look at different aspects of the design and develop a qualitative and quantitative profile that characterizes the design in practice. In our study we focus on the development and application of criteria in relation to ways of students' reflection. Another characteristic of design research is the contribution of different participants to the design, in order to bring their different sorts of expertise into the production and analysis (ibid, p. 22). In our study teachers were explicitly involved in the designing and refining of the project.

Our design research project was carried out in two research cycles (cf. Boersma et al. 2005; Schalk 2006). In this article we describe part of the second research cycle.

Samples

Our study was carried out in two schools for pre-university education, one school in Amsterdam, and the other school in Amstelveen. Students from the school in Amsterdam mostly came from upper middle class families, whereas the students from the school in Amstelveen came from all kinds of social classes. In each school a teacher with more than 5 years of experience in biology teaching instructed the students. The students were in the 11th grade, age 16 or 17, with a total of 50 students. Besides the subject of biology, nearly all students in our study also did mathematics, physics and chemistry. All students had some experience in doing investigations, although not of the open-ended character as in this study. In earlier projects, some aspects of the quality of investigations had been addressed, like the formulation of research questions, the testing of hypotheses, the accuracy of measurements, the link between the research question and the conclusion, and the reliability of an investigation (Schalk 2006).

In each school one group of students followed the ROA strategy and another group followed the RIA strategy. The lessons of the different groups were not at the same time, so the teachers could give each group their full attention.

The allocation of the students to the groups was done by their teachers, after a request from the researchers to construct two groups as equal as possible with regard to achievement and interest in biology. Within each ROA and RIA group, two couples of students (in the RIA group in Amsterdam one couple and one trio) were selected to be studied more closely by audio-taping all their conversations in class, making a total of eight small groups including 17 students. These students were selected by the

Table 2 Size, composition and achievement of the groups of students in the classes in Amsterdam and Amstelveen and of the selected small groups

All students	Reflection-on-action	Reflection-in-action	Average pretest score (% of maximum)
School in Amsterdam	10 (4♀, 6♂)	11 (6♀, 5♂)	53
School in Amstelveen	15 (9♀, 6♂)	14 (13♀, 1♂)	52
average pretest score (% of maximum)	52	51	
Selected students			
School in Amsterdam	4 (2♀, 2♂)	5 (3♀, 2♂)	54
School in Amstelveen	4 (2♀, 2♂)	4 (3♀, 1♂)	52
average pretest score (% of maximum)	52	54	

teacher and were average with regard to achievement and interest compared to the classes as a whole. The achievements on the pre-test -which is referred to earlier, see Schalk et al. (2009)- were comparable for the classes, the RIA and ROA groups as well as for the selected small groups (see Table 2).

The Inquiry Project

How does the above described theoretical framework lead to the design of teaching and learning strategies for learning to ensure the quality of investigations? We derived the following design principles.

- The teaching and learning strategies are focused on procedural understanding, i.e. knowledge about and understanding of the quality of investigations, not on substantive knowledge and understanding; the conceptual content is formed by the knowledge base of the concepts of evidence (Millar et al. 1994; Séré 2002*)*.
- To stimulate motivation for learning about the quality of investigations, students can choose their own subject and research question: what do they really want to know? This 'ownership' is known to enhance students' involvement (Dudley-Marling and Searle 1995).
- Students carry out an investigation of their own, in which they are allowed to make most choices by themselves. As Gott and Duggan (1995b, p. 144) state: In extended, open-ended investigations students ". . . will be required to synthesize all of these ideas into an overall solution". Here, 'ideas' stand for the concepts of evidence.
- There are many opportunities for interaction: students work in small groups of two or three and the teacher is available for questions and feedback.
- The acquisition of knowledge is stimulated by reflection on the quality of their own investigations; this is done by appealing to what they already know and can do.

The inquiry project was carried out in two versions: (1) using the *reflection-in-action (RIA)* strategy and (2) using the *reflection-on-action (ROA)* strategy.

The nine-lesson project started in both versions with an introductory lesson to provide a common basis about what should be understood by 'good investigations' based upon what the students already knew. Most of the lesson was filled with a whole-class discussion about designing an investigation. At the end of that lesson students were provided with a booklet to support their inquiry project. The booklet contained a description of various types of inquiry and a selection of the criteria to evaluate the quality of an investigation, based on the concepts of evidence. The reason to limit the number of criteria lies in the fact that addressing 23 criteria in one investigation project is not feasible for both the teacher and the students. The selection of criteria was made in collaboration with a forum of experienced teachers, including the teachers in the project. Criteria were selected about how to formulate a good research question (criteria nos. 1 and 18, Table 1), about hypothesis testing (nos. 3, 4, 5 and 6), about controlling variables and taking samples (nos. 8, 9, 10 and 14), and about reliability and validity (21). These selected concepts constitute the core of the hypothetical-deductive method used in most biology investigations (Allen and Baker 2001), in combination with an emphasis on reliability and validity (Gott et al. n.d.).

In both strategies students carried out an investigation in small groups of two or three. The small groups formulated their own research questions and made their own research designs. Because the selection of criteria was centred around the hypothetical-deductive method, students were asked to design a hypothesis testing investigation. The subject of the investigation could be anything within the scope of biology, for instance, the effect of anti-bacterial mouthwash on the amount of bacteria on teeth after different periods of time, the relation between the smoking behaviour of students and that of their parents, or the relation between age and the functioning of short-term memory. During eight lessons students were working on their projects in class. Additionally, they carried out parts of their investigations at other moments: in school, at home or in the field. Only in class could they ask questions of their teacher.

In both strategies, students reported the design and results of their investigations in a ‘scientific’ article, for which guidelines were given in their booklet.

In the *ROA teaching and learning strategy,* reflection-on-action was stimulated by giving reflection tasks to the students. Four times, students were temporarily diverted from their own investigations, but afterwards they were prompted to reflect on their own investigations with the criteria that resulted from the task given. All selected criteria were addressed. Each task presented a problem at the concrete level, asked students to give a solution and to explain why it was a solution, and contained a question about which general criteria for good investigations could be derived from the given problem and its solution. Students carried out the tasks together talking about it in their small groups. Afterwards, every group discussed the task with the teacher.

The first task was about how to formulate a good research question (concepts nos. 1 and 18): three research questions were presented to the students and they were asked which is best and why. From their answers criteria concerning the formulation of research questions were derived and students were asked to reflect on their own research question using these criteria. The second task was about hypothesis testing (concepts nos. 3, 4, 5 and 6), and the third about controlling variables and taking samples (concepts nos. 8, 9, 10 and 14). In the fourth task students were invited to give oral comments on draft versions of each other’s research reports, by using the criteria (especially concept no. 21).

In the *RIA teaching and learning strategy*, questions about how to conduct a good investigation were raised – either by the teacher or by the students – when the quality of the ongoing investigations of the students themselves asked for it, for example: “What do you mean by ‘prediction’?”, “So, how do you measure reproduction [of bacteria]?” or “How certain are you about your conclusion and why?” In the RIA teaching and learning strategy, the students also received written feedback about their own investigations twice. The first time they received feedback on their research plans, the second time on their draft reports. In these comments the teachers incorporated remarks related to the criteria, especially to the selected ones, which were the same as those addressed in the reflection tasks. For example: “What are the dependent and the independent variables in your investigation?” or “Why have you measured this amount twice?”

The assignment of carrying out an open-ended investigation like this draws heavily on the students’ abilities. They have to choose a subject in the scope of biology of which they must have developed enough substantive knowledge to come up with a sensible research question; they must design an investigation, so they should have acquired a range of possible research methods and the skills to perform these methods; they are expected to be curious and display a scientific attitude; and they should have developed enough procedural understanding to ensure the quality of the investigation, in parts and as a whole. The fact that our interest is focused on the development of procedural understanding does not mean that the other factors are of no importance, on the contrary. However, this complex demand on students’ abilities follows from our choice for open-ended investigations about a student-chosen subject in order to provide opportunities to “synthesize all of these ideas (about evidence, [author]) together into an overall solution" (Gott and Duggan 1995b, p. 144). Furthermore, since the substantive demand of an investigative task does not seem to affect the contribution of procedural knowledge (Glaesser et al. 2009a), we assume that the demand for other abilities will not greatly influence the ways of developing procedural understanding.

Instruction of the Teachers

Both teachers participated in the first and the second research cycle. Part of the evaluation of the first cycle was an evaluation with the teachers of the teacher behaviour. After this evaluation both teachers had –according to their reactions– a clear understanding of the ROA and the RIA strategy and the expected teacher behaviour. That behaviour can be characterized as: stimulating reflection-in-action during the project by discussing the students' investigations, either by initiating discussion themselves or in reaction to students' questions; and focusing in conversations on the criteria for good investigations by asking questions about criteria explicitly.

Methods of Scoring and Analyzing

All conversations in class –with or without the teacher participating– of the eight small groups (17 students) were audio-taped and transcribed. The conversations were divided into episodes on the basis of subject or participants, for instance, a new episode started when students shifted their conversation from the design of their investigation to the way to get the teacher's attention, or when the teacher joined the group.

All episodes were divided into utterances, as units of speech in spoken language, which is our unit of analysis. Utterances were bounded on the basis of who was speaking and what he or she was talking about. When one speaker was interrupted by another speaker, both parts of the interrupted utterance were counted as one (e.g. utterance 65 in Table 3), whereas when a speaker was talking for a long time (e.g. the teacher), or was changing, for instance, from description to explanation, this fragment of speech was divided into several utterances. The methodology is qualitative and primarily focused on the semantic content, but also took into account the relationships among language and context (Lemke 1998). The fact that the utterances are written down as if only one person at a time was speaking does not fully mirror reality, nor does the fact that most non-verbal communication is left out. Nevertheless, we agree with Ochs (1979) that it is adequate for adult-adult conversations in which most utterances are related to the previous ones, and with Mortimer and Scott (2003) that talk between teacher and students, or among students, shares this adult feature.

All utterances of students and teachers were analyzed by two independent researchers. The first distinction was whether the utterances (in their verbal and non-verbal context) had a relation to the quality of investigations or not. All the utterances that had such a relation (3943 in total) were analyzed further with respect to the criteria they were related to. The proportion of the number of utterances about which there was agreement between both researchers was calculated (0.79), as was Cohen's kappa for inter-rater agreement (0.58). After that, the researchers discussed their analyses until consensus was reached. A second distinction of the utterances was whether they contained a *problematization,* a *description,* an *explanation,* a *generalization* or an *application* (proportion agreement = 0.84; Cohens' kappa = 0.64; discussion until consensus). These categories were described as:

Table 3 Scoring of a transcript of a part of an episode of a classroom discussion. In the right hand column three patterns of utterances are mentioned, one in italics, one in bold scores, and one in non-italic, non-bold; the constitutive utterances are also indicated by italics or bold scores. The discussion is about the question of how to investigate whether blond girls are less smart than other girls

Epi	Utt	Speaker	Text	Cat/crit	Pattern
04	47	Teacher	Are there perhaps other things that influence …, except age?	P9	
	48	Student 1	Yes, for instance, it depends on whether you only consider people in pre-university education, or …	D9	
	49	Student 2	Criminal background.	D9	
	50	Teacher	Ah, education!	D9	
	51	Student 3	But education should not be of influence either, actually	P9	
	52	Student 4	[inaudible]		
	53	Teacher	I saw a hand raised over there …		
	54	Student 5	Yes. How long is … is your IQ …?	*P9*	
	55	Teacher	How tall* they are, those girls? No.	D9	
	56	Student 4	Your brain size?	P9	
	57	Teacher	That kind of things. But you wanted to ask something?	D9	
	58	Student 5	But an IQ-test, once it is administered, is it for the rest of your life, so to speak …?	*P9*	
	59	Teacher	Martin says it is.	*D9*	
	60	Student 2	Yes, it is not …		
	61	Student 5	I mean …, does it always stay the same, doesn't it change in the course of time, in the course of years?	*P9*	
	62	Teacher	An IQ-test actually measures what you are capable of and not how much you know or something, like that.	*E9*	*PDPE9* ► *DE9*
	63	Teacher	Ehm, well, we are a little stuck then with that problem. ain't we? Then, ehm, what education should those girls have or doesn't that matter?	P9	
	64	Student 2	No, because you will notice that, right, no. No.	D9	
	65	Teacher	I don't know, just tell me. How …	P9	
	66	Student 6	No.		
	(65)	Teacher	… how are you going to put together those groups, how will they look like? Now for [student 6].	**P14**	
	67	Student 6	The best will be if [inaudible] [probably student 6 says that it would be best to take twins, one of which is blond and one is not].	D9	
	68	Student 5	That won't go.	P9	
	69	Teacher	Well, it can. But it is just seldom occurs, indeed.	D9	
	70	Student 2	Mutate!		
	71	Student 6	Well, a twin, but non-identical, then. That are the same [inaudible]	E9	
	72	Teacher	Okay, so you say …, So, you say, best would be if you – because a group of one is a little small – if you have bin ovular twins, you split them in two, and you carry out your investigation with them. That is possible, but it would be nicer if you would have … a hundred of them.	**D14**	
	73	Student 6	But two …		
	74	Teacher	But what would those groups look like? Unfortunately, I think a bin ovular twin will be difficult.	**P14**	**PDP14** **►D14**

Table 3 (continued)

Epi	Utt	Speaker	Text	Cat/crit	Pattern
	(73)	Student 6	… bin ovular twins also just have the same background, same … education …	E9	
	75	Teacher	So you say, ehm, okay, the members of both groups, the blond and the other one, should correspond, in everything, age, background ..	G9	$[PD]_6$EG9 ►DEG9

Epi episode, *utt* utterance, *cat* categorization [in which *P* problematization, *D* description, *E* explanation, *G* generalization, *A* application], *crit* criterion [where the number refers to the list in Table 1]. *: in Dutch, the word for 'long' and 'tall' is the same

- *Problematization (P)*: a student or a teacher poses a question or signals a problem for which an answer or solution is required.
- *Description (D)*: a student or a teacher describes a possible answer or solution to a specific problem.
- *Explanation (E)*: a student or a teacher explains what the idea or principle behind a specific answer or solution is.
- *Generalization (G)*: a student or a teacher formulates a general criterion or rule.
- *Application (A)*: a specific or general answer or solution is applied to a new problem or in a new context.

The percentages of utterances in these categories in conversations in different situations (e.g. RIA versus ROA) were analyzed for significant difference using the 95% confidence interval: 1.96 times the square root of the sum of the squares of the standard errors, see e.g. Witte and Witte (2004).

In Table 3 a part of an episode is shown to illustrate the method of analysis. It is a part of the classroom discussion in the first lesson about how to investigate whether blond girls are less smart than other girls. The fragment starts with a question from the teacher for more influencing factors – that one should keep all factors constant or kept under control, even though the specific criterion (no. 9, see Table 1) has not been mentioned explicitly – except the age of the girls which has been mentioned earlier in the episode. This problematization (scored as P9) is followed by the description of some concrete factors mentioned by students (utterances 48, 49 and 56, scored as D9) and asserted by the teacher (50, 57, also scored as D9). Student 3, in his turn, problematizes these factors (51, P9). Then, suddenly, another issue arises. The question of student 5 (54) is at first misunderstood by the teacher (55), but after the student reformulates the problem (58), the teacher understands what it is about, i.e. whether IQ is age-dependent. This can also be interpreted as a reference to concept no. 9, because it is about the factors influencing the testing of one's IQ. This question reverts to a discussion that took place prior to this fragment. Therefore, the teacher refers to something that has been said earlier by Martin (59); as student 5 insists, the teacher explains the idea behind an IQ-test (62, scored as E9). The fact that this is incorrect (IQ can change) does not matter for the scoring. Then, the teacher picks up the previous line of discussion (63, P9) and broadens it to the composition of a sample (65, scored as P14). Student 6 then comes with a suggestion to control the influence of other factors: use twins, one of which is blond and the other is not (67, she speaks rather softly and what she says exactly cannot be heard on the recording; however, from the continuation can be derived what she probably said). As student 5 problematizes her suggestion (68, P9), the teacher contradicts it and student 6 explains what she means (71, E9). The teacher picks up the suggestion, but problematizes whether you can get enough of those twins (72, P9). Interrupted by the teacher, who switches to the question of sample composition (73, P14), student 6 explains it again and more extendedly (74, E9), which is then picked up and generalized by the teacher (75, scored as G9).

In a series of utterances patterns can be observed. If in an episode, a concept of evidence is discussed on the descriptive, the explanatory as well as the general level, we call it a *complete pattern*. In Table 3, the discussion about influencing factors is such a pattern. If only two of those three levels occur, we call it an *incomplete pattern*, for example, the discussion about the IQ-test in Table 3 (indicated by scores in italics that constitute the pattern). A *rudimentary pattern* includes only one of those three levels, like the utterances about sample composition which (indicated by bold scores). The category 'problematization' is not incorporated in the scoring of the patterns, because this can occur at any level. And in our scoring, the presence of application is not necessary for a complete pattern. If more than one utterance in a certain category is observed in a pattern about a certain concept of evidence, this is ignored. So, the rudimentary pattern of PDP14 is scored as D14, the incomplete pattern of PDPE9 is scored as DE9, and the complete pattern of [PD]$_6$EG9 is scored as DEG9. Table 4 lists which patterns of utterances are scored in these categories.

Findings

The first remark can be that we succeeded in categorizing utterances with respect to the criteria and the levels in Fig. 3, with sufficient inter-rater reliability.

The results of the categorization of all the utterances of students and teachers working in small groups are shown in Fig. 4. It can be seen that some criteria were more often the subject of discussion than others. Especially the criteria concerning the research question (no. 1, see Table 1) and the variables (nos. 8 and 9) were often addressed; the concepts concerning the hypothesis (nos. 3 and 4) and the ways of measuring (nos. 12 and 13) were popular as well. The overarching criterion no. 21 was only seldom discussed. This observation holds as well for the conversations about the students' investigations (Fig. 4, top) as well as for the conversations concerning the reflection tasks (Fig. 4, bottom). So, not surprising, most conversations were about the criteria that were selected to be explicitly addressed in the project –as emphasized in the teachers' written feedback and the reflection tasks- but other criteria were discussed as well and some criteria, although included in the selection, were not very often discussed.

Table 4 Scoring categories of complete, incomplete and rudimentary patterns

Complete patterns	Incomplete patterns		Rudimentary patterns
DEGA	DEA	GEA	D
DEG	DE	GE	E
GEDA	DGA	GDA	G
GED	DG	GD	
otherwise complete	EGA	EDA	
	EG	ED	
	otherwise incomplete		

D description, *E* explanation, *G* generalization, *A* application

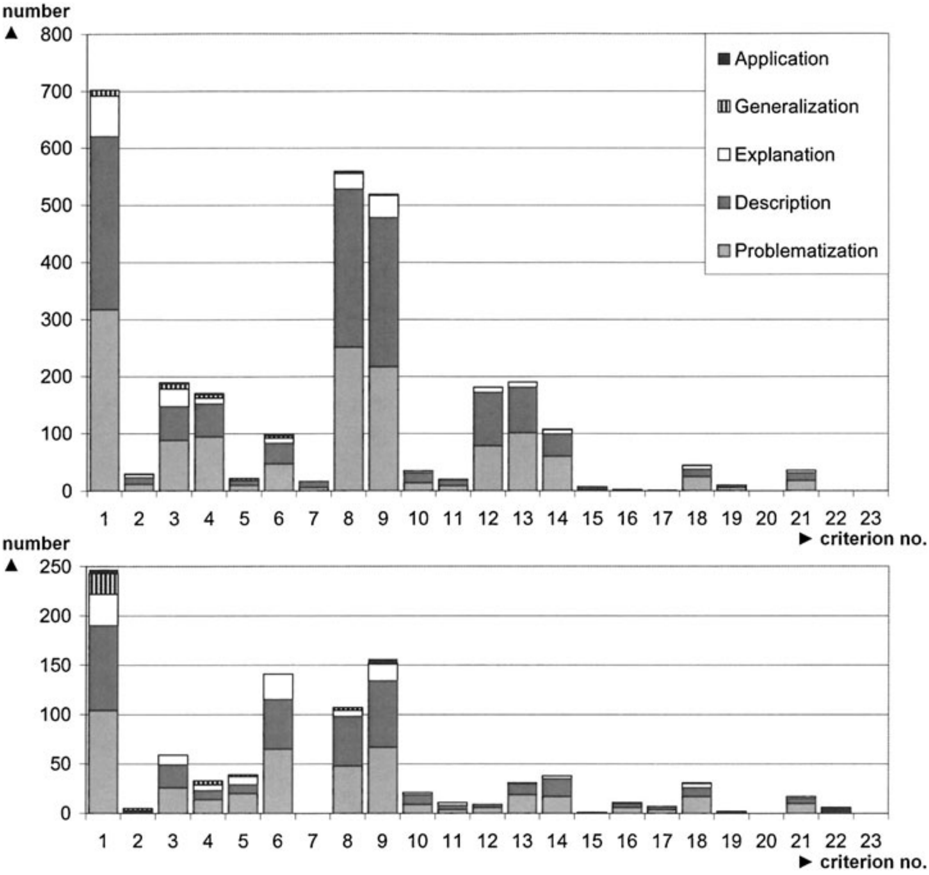

Fig. 4 Numbers of utterances (of all students and teachers together) about different criteria for evidence (cf. Table 1) in the categories Problematization, Description, Explanation, Generalization and Application. Top: utterances concerning students' own investigations; bottom: utterances concerning the reflection tasks

The distribution of the utterances over the five categories *problematization, description, explanation, generalization* and *application* is also shown in Fig. 4. It can be seen that the majority of the utterances are at the concrete level of problematizing and describing. In the conversations about the reflection tasks the contribution of explanatory utterances is higher, which is not surprising, because these tasks explicitly ask for explanation and generalization.

An important question is by whom the different types of utterances were made. This is shown in Fig. 5. The majority of the teachers' utterances were problematizing (Table 5 and Fig. 5, outer circles). This means the teachers did what they were supposed to do: asking the students questions to stimulate them to describe, explain and generalize. The students did give answers (middle circles), but most of these were at the descriptive level. Most of the explanations, generalizations and applications came from the teachers (17.8% of their total [117 utterances] to 7.6% [57] of the students). In mutual conversations between students also, relatively few utterances at the explanatory or general level were observed (inner circles, 6.8%, 103 utterances). So, these conversations also remained merely at the concrete level of problematizing and describing the ins and outs of the investigations. For instance, one couple discussed how many people they should interrogate about smoking behaviour without referring

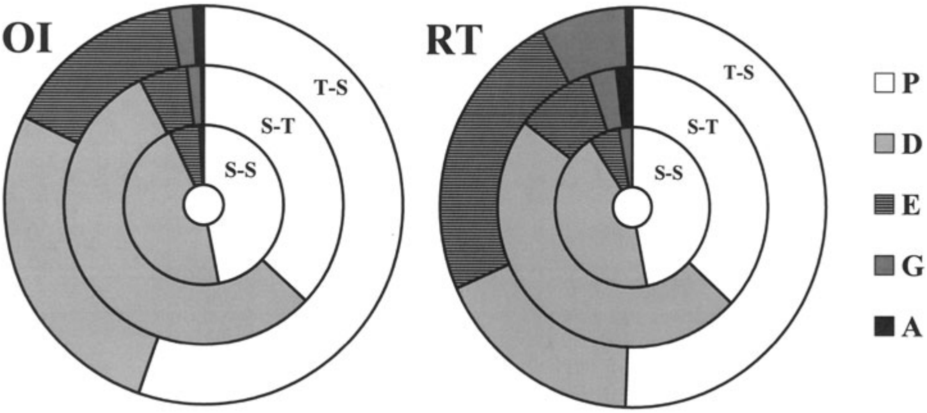

Fig. 5 Categorization of utterances of teachers talking to students (T-S, outer circles), of students talking to teachers (S-T, middle circles), and of students talking to students (S-S, inner circles) in conversations about their own investigations (OI, left, 2922 utterances) and in conversations about reflection tasks (RT, right, 1021 utterances). P: problematization; D: description; E: explanation; G: generalization; A: application

to the concept of a representative sample, but only thinking of practicability, i.e. how much time they had and how many 11th graders there were in their school.

As already noted, in the conversations about the reflection tasks which were constructed to stimulate explanation and generalization, the percentages of utterances in the categories explanation, generalization and application were significantly higher (Fig. 5, right, and Table 5).

If the data of the conversations about students' own investigations are divided into the ROA and the RIA strategies (Fig. 6 and Table 6), it can be seen that the teachers did not make more explanations, generalizations or applications in one strategy compared to the other, but the students who had done the reflection tasks (i.e. the ROA strategy) made significantly more explanatory and generalizing utterances in their conversations with the teachers about their own investigations, (Fig. 6, middle respectively outer

Table 5 Percentages of utterances in the categories problematization, description, explanation, generalization, and application in conversations about students' own investigations or about the reflection tasks

	Teacher to student			Student to teacher			Student to student		
Category	p_{OI} (%)	p_{RT} (%)	p_{RT} -p_{OI}	p_{OI} (%)	p_{RT} (%)	p_{RT} -p_{OI}	p_{OI} (%)	p_{RT} (%)	p_{RT} -p_{OI}
P	55.3	50.6	−4.6	37.0	37.0	−0.1	46.9	47.0	0.1
D	26.9	17.7	−9.1*	55.4	48.2	−7.2*	46.3	43.8	−2.5
E	15.1	23.9	8.8*	5.7	9.6	4.0*	5.9	6.5	0.6
G	1.8	7.1	5.3*	1.5	3.2	1.8	0.7	2.8	2.0*
A	0.9	0.6	−0.3	0.4	1.9	1.5	0.2	0.0	−0.2
E + G + A	17.9	31.6	13.8*	7.5	14.8	7.3*	6.8	9.3	2.4
N	1511	400		756	311		655	310	

P problematization, *D* description, *E* explanation, *G* generalization, *A* application, *N* number of utterances, p_{OI} percentage in conversations about students' own investigations, p_{RT} percentage in conversations about the reflection tasks; *: $p<0.05$ (significance: the absolute difference is greater than the 95% reliability interval, calculated by 1.96 * sqrt{p_{OI} * [100-p_{OI}]/[N_{OI}-1] + p_{RT} * [100-p_{RT}]/[N_{RT}-1]})

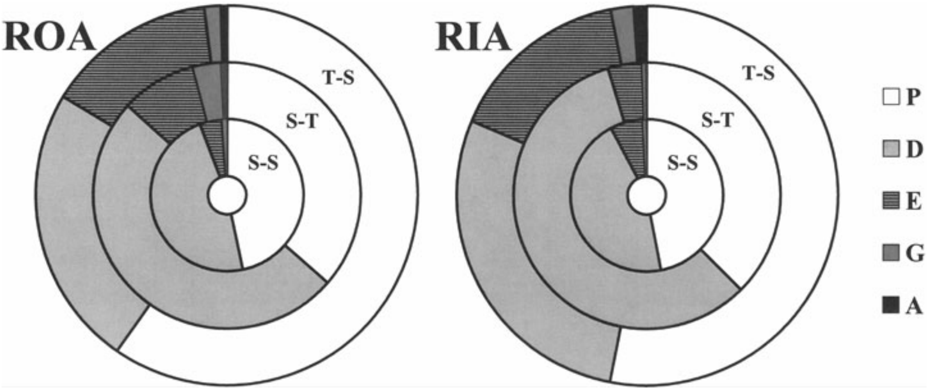

Fig. 6 Categorization of utterances of teachers talking to students (T-S, outer circles), of students talking to teachers (S-T, middle circles), and of students talking to students (S-S, inner circles) in conversations about their own investigations. Left: students from the group which followed the reflection-on-action strategy (ROA, 1658 utterances); Right: students from the group which followed the reflection-in-action strategy (RIA, 1264 utterances). P: problematization; D: description; E: explanation; G: generalization; A: application

circles and Table 6). In students' mutual conversations, no significant difference was observed between the two teaching and learning strategies (Fig. 6, inner circles and Table 6).

The observed patterns of utterances in the different categories reflect the emphasis on the descriptive level (Fig. 7). Only very few complete (consisting of D, E and G) and a minority of incomplete (consisting of two of these three categories, see Table 4) patterns were observed; the majority consisted of rudimentary (consisting of only one category) patterns. The majority of the incomplete patterns and nearly all complete patterns occur in conversations between teacher and students, and nearly no complete patterns were observed in the conversations between students.

Table 6 Percentages of utterances in the categories problematization, description, explanation, generalization, and application in conversations about students' own investigations in both strategies

	Teacher to student			Student to teacher			Student to student		
Category	p_{ROA} (%)	p_{RIA} (%)	p_{RIA} - p_{ROA}	p_{ROA} (%)	p_{RIA} (%)	p_{RIA} - p_{ROA}	p_{ROA} (%)	p_{RIA} (%)	p_{RIA} - p_{ROA}
P	59.6	53.2	−6.5	36.3	37.4	1.1	46.8	47.1	0.3
D	23.9	28.3	4.3	50.4	58.0	7.6*	47.3	45.1	−2.2
E	14.6	15.4	0.8	9.0	4.0	−5.0*	4.9	7.0	2.1
G	1.4	2.0	0.6	3.5	0.4	−3.1*	0.9	0.6	−0.3
A	0.5	1.1	0.7	0.8	0.2	−0.6	0.1	0.3	0.2
E + G + A	16.4	18.6	2.1	13.3	4.6	8.7*	5.9	7.8	1.9
N	213	442		256	500		795	716	

P problematization, *D* description, *E* explanation, *G* generalization, *A* application, *N* number of utterances, p_{ROA} percentage in the reflection-on-action strategy, p_{RIA} percentage in the reflection-in-action strategy; *: $p<0.05$ (significance: the absolute difference is greater than the 95% reliability interval, calculated by 1.96 * sqrt{p_{ROA} * [100-p_{ROA}]/[N_{ROA}-1] + p_{RIA} * [100-p_{RIA}]/[N_{RIA}-1]})

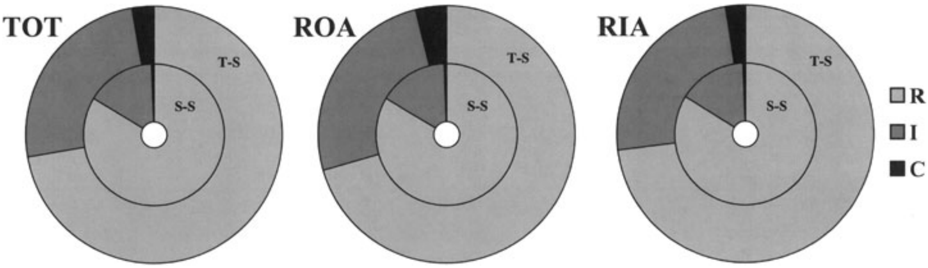

Fig. 7 Categorization of patterns of utterances in conversations about students' own investigations between teacher and students (T-S, outer circles) and among students (S-S, inner circles). Left: all conversations (TOT, 672 patterns); centre: conversations in the reflection-on-action strategy (ROA, 287 patterns); right: conversations in the reflection-in-action strategy (RIA, 385 patterns). R: rudimentary patterns; I: incomplete patterns; C: complete patterns

Conclusions and Implications

Describing, Explaining and Generalizing

The first research question was how pre-university students develop procedural understanding, i.e. how they acquire criteria for the evaluation of the quality of investigations, and whether the spiral of description, explanation and generalization could be observed in the practice of students carrying out an inquiry project in biology. Although we succeeded in categorizing the utterances, complete patterns of reasoning in the utterances that include description, explanation as well as generalization were not often found in the conversations of the 17 students we observed. Especially in their mutual conversations students hardly surpassed the level of description. Furthermore, in conversations in which the teacher was present, most of the explanations and generalizations were not made by the students but by the teacher. Yet, the majority of the utterances of the teacher falls in the category of problematizing. So, even in a dialogic discourse (Mortimer and Scott 2003), where teacher and students can both contribute to meaning making, the students rarely enter the higher levels. The near absence of utterances on the explanatory and general level is reflected in a majority of rudimentary patterns: instead of a spiral movement that leads to higher levels of understanding, students swing to and fro between problematizing and describing at the concrete level of their own investigations. In general, only when the teacher participates explanations and generalizations enter the conversations. Consequently, the conclusion is that the 'spiral formation of mental actions' could only be observed in the presence of the teacher. Only then the complete patterns occurred in which description as well as explanation and generalization are interconnected. Therefore, his or her presence appears to be indispensable and crucial for stimulating the development of procedural understanding. This means that the social plane on which meaning making takes place should include the teacher: he or she is making the 'scientific story' available (Ogborn et al. 1996; Mortimer and Scott 2003).

Does this conclusion mean that 'doing science' is not the right way or is not enough to realize the internalization of the criteria for evidence? This would be in line with the statement by Lederman cited in our introductory section. But we tried, by explicitly stimulating reflection on criteria, that what our students did, would be more than just 'doing science'. Then, does it mean that a more 'instructive' approach is needed to get the criteria across? We think this leads to implications for the teacher's role in the guidance of students' investigations, which will be discussed in a following section.

The Role of Reflection

The second research question was whether a difference could be observed between the two teaching and learning strategies with regard to stimulating explanations and generalizations.

It appeared that the reflection tasks in the ROA strategy on their own did not stimulate students to explain and generalize, but that the conversations with the teacher about the tasks did. And they did significantly more than the conversations about the investigations of the students. Still, even then the majority of the explanations and generalizations came from the teacher. However, in their conversations about their own investigations, students who did the reflection tasks explained and generalized significantly more than students who did not do the tasks. This might be explained by the fact that the teacher has a 'point of reference', as can be seen in the following transcript, in which he addresses the criterion "hypothesis should be testable" (criterion no. 4, Table 1). The research question of the student is whether boys can type faster than girls, with a possible explanation that boys spend more time behind the computer than girls. The teacher refers to reflection task 2, which is about behaviour of gazelles.

Teacher:	My question, with which I started, was "Is it testable?"	P4
Student:	Yes.	
Teacher:	How do you know?	
Student:	What do you exactly mean by testable?	P4
Teacher:	Whether you can test the hypothesis.	P4
Student:	That it is feasible?	P4
Teacher:	That is what I wanted to say …, well, an investigation is feasible, a hypothesis, you test.	E4
Student:	Yes, I think so, because you just time it to see how fast it's all going and you let them fill in how many hours a day or how much time a day they sit behind the computer. And then you start working on it and if it then, if it then shows that boys are much faster. You first only look at the time and after that you look at the number of hours that they then sit behind the computer. So I think it is testable.	E4
Teacher:	Yes, because what in the example with the gazelle made …, what makes that hypothesis testable? At which moment does the researcher know, I can test the hypotheses?	P4
Student:	Here. [*points at the place in reflection task 2 where the observation is described*]	D4
Teacher:	That was afterwards already, this? This is an observation already.	P4
Student:	You mean before, when he makes the hypotheses?	P4
Teacher:	Yes. Before he goes into the field.	
Student:	He has three clear, three clear differences, that can be observed, so to speak, if you are going to investigate that.	D4
Teacher:	Yes.	
Student:	So you can think, that is testable.	E4
Teacher:	Yes, that is it. He has, hasn't he, … these are actually predictions he does here in case the hypothesis is correct.	E4
Student:	Yes.	
Teacher:	So that, at the moment you have that, and when there is a clear distinction in it of course, because yes, … then you know in any case that your hypothesis meets the criterion, it is testable.	G4

We can now reconsider the dilemmas related to the ways of reflection we encountered when designing the teaching and learning strategies. The dilemma of content –should the reflection be restricted to the criteria that students run into themselves?– tends towards a 'no'. The teachers asked questions on any of the selected criteria in relation to the students' investigations, especially in the written feedback, and also in the cases students themselves did not problematize. The dilemma of initiative –do you wait to reflect until students question the quality of their investigations themselves?– was not a dilemma for the designers; it should be done and it was done by the teacher. But it deserves a remark here that the written feedback of the teachers in the RIA strategy and the fourth reflection task, where students gave oral comments to their peers, was highly appreciated by the students. The dilemma of moment –should the teacher interrupt students' investigations for moments of reflection?– tends towards a 'yes'. Taking a side-step from one's own investigations seems to stimulate the spiral of description, explanation and generalization more than reflection-in-action alone. Put otherwise, recontextualizing (Van Oers 1998) the criteria in an additional context seems to stimulate the formation of mental actions a little more than staying in the context of students' own investigations. It will require further research to find out which conditions will favour the subsequent abstraction best (Van Oers 2001).

The Design of the Project

As we reported elsewhere (Schalk et al. 2009) and summarized earlier, both strategies yield enough learning effects to be 'good enough' to teach and learn understanding of evidence, the RIA strategy a bit more in the students' reports, the ROA strategy a bit more in the evaluation of others' investigations. Then what do the conclusions formulated above mean for the design of the project, as it might be taught in the future? Should it have reflection tasks to promote reflection-on-action or should it just stimulate reflection-in-action? On the basis of our findings we would suggest a combination of the strengths of both teaching and learning strategies: written feedback on students' own investigations together with side-steps to reflect on the criteria for good investigations.

Additionally, the introduction of the criteria could be strengthened. In our design, the criteria are derived in a whole class discussion about an imaginary investigation in the first lesson. In recent research, Roberts et al. (2010) have shown the effects of explicitly teaching the knowledgde base of procedural understanding on the performance in open-ended investigations to be substantial. They suggest it might limit the number of investigative tasks necessary to develop procedural understanding. In our design, such teaching might function as a point of reference for discussions about the quality of investigations, as do the reflection tasks.

Nevertheless, we would suggest that more than just one inquiry project should be done. Although it needs to be said that whatever the time span –nine lessons in 4 to 5 weeks– might be the maximum possible in many schools, it is still a short time to develop thorough procedural understanding. As we stated earlier, the project is a complex task in which students have to give attention to the substantive knowledge, technical affairs and planning in addition to the reflection on the quality of their investigations. Taking this into account, the limited progression of procedural understanding can be valued positively. Making the inquiry project part of a learning trajectory that spans all years of secondary education and addresses every criterion more than once or twice, as suggested by others (Van der Valk and Van Soest 2004), may increase expectations for procedural understanding. In fact, we now are working on a more extended learning trajectory in the course of pre-university education, and not only in biology, but in chemistry and physics as well.

The Teacher's Role

Although it was not the focus of our study, the role of the teacher prominently surfaces. The process of generalization nearly only takes place in the teachers' presence, this being –in our model– indispensable for internalization of the criteria, the intended learning outcome. The role of 'cranking up' this process is an extension to the roles Crawford (2000) attributes to teachers in inquiry classrooms (motivator, diagnostician, guide, innovator, experimenter, researcher, modeller, collaborator and learner). It goes beyond the role as a diagnostician giving students opportunities to express their ideas, or the role as a guide, helping students to develop strategies (the example Crawford gives lies, in our view, on the descriptive level). This role includes explicitly entering the higher levels of explanation and generalization. It approaches the pedagogical intervention Mortimer and Scott (2000, p. 132) call "developing the epistemological line," "introducing students to the nature of scientific knowledge (such as the generalizability of scientific explanations) which is being taught," but focuses more on the development of criteria students can use for it. It also is an extension to the role of modeller "showing the attitudes and attributes of scientists by example" (Crawford 2000, p. 932), because the specific attitude of applying criteria for evidence has to be made explicit.

And how should this role be carried out? In our study, teachers asked the students questions to stimulate their thinking about the criteria. But even then, seldom all levels of describing, explaining and generalizing were passed through. Since the teachers' role was not the focus of our study, we did not categorize the teachers' questions, which might have been useful. On the basis of others' publications, though, some things can be said about it. In a chapter on questioning and discussion in inquiry-based teaching in a book for student science teachers, Bybee et al. (2008, pp. 240–261) recommend divergent questioning and a wait time of more than 3 seconds to stimulate critical thinking. Van Zee (2000) also promotes attentive silence and reticence to give students an opportunity to formulate answers. We think, however, that it requires a more active role in making the path from descriptions to explanations and generalizations explicit, by posing thought stimulating, but directed questions or by articulating conclusions. After a divergent discussion, the teacher should try to converge it to the 'scientific story' of the criteria for evidence by explicitly asking for or stating him/herself a generalization. He or she can do that for instance by referring to an example in another investigation, to a reflection task that was carried out by the students or to earlier lessons about the criteria. Then, the transfer of ideas about evidence from teaching to their application in open-ended investigations, as Roberts et al. (2010, p. 404) suggest happens more or less unconsciously, might be made explicit and stimulated.

Just like Krajcik et al. (1998, p. 348) found that middle school "teachers' suggestions and questions proved crucial in encouraging students to be thoughtful about the substantive aspects of their investigations", this study indicates that high school teachers (should) play a similar 'pivotal role' (Kelly 2007, p. 451–452) in encouraging students to be thoughtful about the criteria for good investigations. 'Pivotal' in the sense that the teacher (re)directs the discussion in the desired way, i.e. towards generalization of criteria.

How can teachers learn to do this? First of all, by conducting investigations themselves, because the teachers' experience with scientific investigations strongly influences the way teachers support students' investigations (Windschitl 2003, 2004). Then, teachers can learn to scaffold students in carrying out open-ended inquiries when they are scaffolded themselves (Van der Valk and De Jong 2009). In our view the scaffolding tools could be made more explicit by having the teachers focus on the criteria for good research.

However, before drawing too big a conclusion from a small sample, further research should be –and will be– conducted to determine ways in which teachers can (learn to) stimulate generalizations of the criteria for evidence.

Implications for Teaching

We conclude this article with a summary of the implications for teaching understanding of evidence that, in our view, come from our findings.

- Let students carry out open-ended investigations more than once. Let them apply the criteria for evidence in more than one context. Let students reflect on the quality during their own investigations *–reflection-in-action– and* by taking a step aside *–reflection-on-action–* not only afterwards, but also during the investigation itself.
- In conversations about investigations of students, stimulate talking on the explanatory and general level of the criteria for evidence. Explicitly ask questions to force students to give an explanation or to formulate a general concept.
- Stimulate complete patterns of reasoning, either by stating a general criterion and asking where this can be observed in the investigations of the students, or by extrapolating utterances on the concrete, descriptive level to the explanatory and general level. Ask students to apply the criterion to another (imaginary) investigation.
- In conversations about the students' own investigations, refer to the criteria for evidence as seen in previous investigations or reflection tasks.

References

Abd-El-Khalick, F., BouJaoude, S., Duschl, R., Lederman, N. G., Mamlok-Naaman, R., Hofstein, A., Niaz, M., Treagust, D., & Tuan, H.-L. (2004). Inquiry in science education: international perspectives. *Science Education, 88*(3), 397–419.

Allen, G. E., & Baker, J. J. W. (2001). *Biology. Scientific process and social issues*. Bethesda: Fitzgerald Science.

Arievitch, I. M., & Haenen, J. P. P. (2005). Connecting sociocultural theory and educational practice: Galperin's approach. *Educational Psychologist, 40*(3), 155–165.

Bailin, S. (2002). Critical Thinking and Science Education. *Science, & Education, 11*, 361–375.

Boersma, K., Knippels, M.-C., & Waarlo, A. J. (2005). Developmental research: The improvement of learning and teaching of science topics. In J. Bennett, J. Holman, R. Millar, & D. Waddington (Eds.), *Making a difference. Evaluation as a tool for improving science education* (pp. 85–98). Münster: Waxmann.

Brown, A. L. (1992). Design experiments: theoretical and methodological challenges in creating complex interventions in classroom settings. *Journal of the Learning Sciences, 2*(2), 141–178.

Buffler, A., Allie, S., Lubben, F., & Campbell, B. (2001). The development of first year physics students' ideas about measurement in terms of point and set paradigms. *International Journal of Science Education, 23*(11), 1137–1156.

Bybee, R. W., Carlson Powell, J., & Trowbridge, L. W. (2008). *Teaching secondary school science: Strategies for developing scientific literacy* (9th ed.). Upper Saddle River: Prentice Hall.

Chin, C., & Brown, D. E. (2000). Learning in science: a comparison of deep and surface approaches. *Journal of Research in Science Teaching, 37*(2), 109–138.

Collins, A., Brown, J. S., & Newman, S. E. (1989). Cognitive apprenticeship: teaching the crafts of reading, writing and mathematics. In L. B. Resnick (Ed.). *Knowing, learning and instruction, essays in honor of Robert Glaser* (pp. 453–494). Hillsdale: Lawrence Erlbaum, Publishers.

Collins, A., Joseph, D., & Bielaczyc, K. (2004). Design research: theoretical and methodological issues. *Journal of the Learning Sciences, 13*(1), 15–42.

Crawford, B. A. (2000). Embracing the essence of inquiry: new roles for science teachers. *Journal of Research in Science Teaching, 37*(9), 916–937.

Dudley-Marling, C., & Searle, D. (1995). *Who owns learning? Questions of autonomy, choice, and control*. Portsmouth: Heinemann.

Duveen, J., Scott, L., & Solomon, J. (1993). Pupils' understanding of science: description of experiments or 'a passion to explain'? *School Science Review, 75*(271), 19–27.

Galperin, P. Ia. (1969). Stages in the development of mental acts. In M. Cole & I. Maltzman (Eds.), *A handbook of contemporary Soviet psychology* (pp. 249–273). New York: Basic Books.

Galperin, P. Ia. (1989). Mental actions as a basis for the formation of thoughts and images. *Soviet Psychology, 27*(2), 45–64. Original work published 1957.

Giere, R. N. (2001). A new framework for teaching scientific reasoning. *Argumentation, 15*(1), 21–33.

Glaesser, J., Gott, R., Roberts, R., & Cooper, B. (2009a). The roles of substantive and procedural understanding in open-ended science investigations: using fuzzy set qualitative comparative analysis to compare two different tasks. *Research in Science Education, 39*, 595–624.

Glaesser, J., Gott, R., Roberts, R., & Cooper, B. (2009b). Underlying success in open-ended investigations in science: using qualitative comparative analysis to identify necessary and sufficient conditions. *Research in Science, & Technological Education, 27*(1), 5–30.

Gott, R., & Duggan, S. (1995a). *Investigative work in the science curriculum*. Buckingham: Open University Press.

Gott, R., & Duggan, S. (1995b). The place of investigations in practical work in the UK NationalCurriculum for Science. *International Journal of Science Education, 17*(2), 137–147.

Gott, R., & Duggan, S. (1996). Practical work: its role in the understanding of evidence in science. *International Journal of Science Education, 18*(7), 791–806.

Gott, R., Duggan, S., Roberts, R., & Hussain, A. (n.d.). *Research into understanding scientific evidence* (pp. 1–14). Durham: School of Education, http://www.dur.ac.uk/rosalyn.roberts/Evidence/cofev.htm, accessed 05-12-2009.

Gott, R., & Murphy, P. (1987). *Assessing investigations at ages 13 and 15.* APU Science Report for Teachers No. 9. London, DES.

Hodson, D. (1993). Re-thinking old ways: towards a more critical approach to practical work in school science. *Studies in Science Education, 22*, 85–142.

Hodson, D. (1998). Mini-special issue: taking practical work beyond laboratory. *International Journal of Science Education, 20*(6), 629–632.

Hofstein, A., & Lunetta, V. N. (2004). The laboratory in science education: foundations for the twenty-first century. *Science Education, 88*(1), 28–54.

Kelly, A. E. (2006). Quality criteria for design research. In J. van den Akker, K. Gravemeijer, S. McKenney, & N. Nieveen (Eds.), *Educational design research* (pp. 107–118). London: Routledge.

Kelly, G. J. (2007). Discourse in science classrooms. In S. K. Abell & N. Lederman (Eds.), *Handbook of research in science education* (pp. 443–469). Mahwah: Lawrence Erlbaum.

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

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

Koschmann, T., Kuutti, K., & Hickman, L. (1998). The concept of breakdown in Heidegger, Leont'ev, and Dewey and its implications for education. *Mind, Culture, and Activity, 5*, 25–41.

Krajcik, J., Blumenfeld, Ph. C., Marx, R. W., Bass, K. M., Fredricks, J., & Soloway, E. (1998). Inquiry in project-based science classrooms: initial attempts by middle school teachers. *Journal of the Learning Sciences, 7*(3&4), 313–350.

Kuhn, D., Amsel, E. D., & O'Loughlin, M. (1988). *The development of scientific thinking skills*. New York: Academic.

Lave, J. (1997). The culture of acquisition and the practice of understanding. In D. Kirshner & J. A. Whitson (Eds.), *Situated cognition: Social, semiotic, and psychological perspectives* (pp. 17–35). Mahwah: Lawrence Erlbaum.

Lunetta, V. N., Hofstein, A., & Clough, M. P. (2007). Learning and teaching in the school science laboratory: An analysis of research, theory, and practice. In S. K. Abell & N. G. Lederman (Eds.), *Handbook of research on science education* (pp. 393–441). Mahwah: Lawrence Erlbaum.

Lemke, J. (1998). Analysing verbal data: Principles, methods and problems. In B. J. Fraser & K. J. Tobin (Eds.), *International handbook of science education* (pp. 1175–1189). Dordrecht: Kluwer.

Lehrer, R., Schauble, L., & Lucas, D. (2008). Supporting development of the epistemology of inquiry. *Cognitive development, 23*(4), 512–529.

Millar, R., Lubben, F., Gott, R., & Duggan, S. (1994). Investigating in the school science laboratory: conceptual and procedural knowledge and their influence. *Research Papers in Education, 9*(2), 207–249.

Mortimer, E. F., & Scott, Ph. H. (2000). Analysing discourse in the science classroom. In R. J. Millar, J. Leach, & J. Osborne (Eds.), *Improving science education: The contribution of research* (pp. 126–142). Buckingham: Open University Press.

Mortimer, E. F., & Scott, Ph. H. (2003). *Meaning making in secondary science classrooms*. Maidenhead: Open University Press.

Ochs, E. (1979). Transcription as theory. In E. Ochs & B. B. Schieffelin (Eds.), *Developmental pragmatics* (pp. 43–72). New York: Academic.

Ogborn, J., Kress, G., Martins, I., & McGillicuddy, K. (1996). *Explaining science in the classroom*. Buckingham: Open University Press.

Reigosa, C., & Jiménez-Aleixandre, M.-P. (2007). Scaffolding problem-solving in the physics and chemistry laboratory: difficulties hindering students' assumption of responsibility. *International Journal of Science Education, 29*, 307–329.

Roberts, R. (2001). Procedural understanding in biology: the 'thinking behind the doing'. *Journal of Biological Education, 35*(3), 113–117.

Roberts, R., & Gott, R. (2002). Investigations: Collecting and using evidence. In D. Sang & V. Wood-Robinson (Eds.), *Teaching secondary scientific enquiry* (pp. 18–49). London: John Murray.

Roberts, R., Gott, R., & Glaesser, J. (2010). Students' approaches to open-ended science investigation: the importance of substantive and procedural understanding. *Research Papers in Education, 25*(4), 377–407.

Rollnick, M., Lubben, F., Lotz, S., & Dlamini, B. (2002). What do underprepared students learn about measurement from introductory laboratory work? *Research in Science Education, 32*, 1–18.

Roth, W.-R. (2009). Radical uncertainty in scientific discovery work. *Science, Technology, & Human Values, 34*(3), 313–336.

Schauble, L. (1996). The development of scientific reasoning in knowledge-rich contexts. *Developmental Psychology, 32*(1), 102–119.

Schalk, H. H. (2006). *Zeker weten? Leren de kwaliteit van biologie-onderzoek te bewaken in 5 vwo.* [Are you certain? Learning to ensure the quality of biology research in pre-university education—with summary in English]. PhD dissertation. Vrije Universiteit, Amsterdam. http://igitur-archive.library.uu.nl/dissertations/2006-1206-200836/index.htm

Schalk, H. H., Van der Schee, J. A., & Boersma, K. Th. (2009). The use of concepts of evidence by students in biology investigations: Development research in pre-university education. In M. Hammann, A. J. Waarlo, & K. Th. Boersma (Eds.), *The nature of research in biological education: Old and new perspectives on theoretical and methodological issues. A selection of papers presented at the VIIth Conference of European Researchers in Didactics of Biology (ERIDOB), Zeist, The Netherlands* (pp. 279–296). Utrecht: Beta Press.

Schauble, L., Glaser, R., Duschl, R. A., Schulze, S., & John, J. (1995). Students' understanding of the objectives and procedures of experimentation in the science classroom. *The Journal of the Learning Sciences, 4*(2), 131–166.

Schön, D. A. (1983). *The reflective practitioner.* Aldershot: Arena, Ashgate Publishing.

Séré, M.-G. (2002). Towards renewed research questions from the outcomes of the European project Labwork in Science Education. *Science Education, 86*(5), 624–644.

Smits, Th, Lijnse, P. L., & Bergen, Th. (2000). Leerlingonderzoek met kwaliteit. [Student research with quality.]. *Tijdschrift voor Didactiek der β-wetenschappen, 17*(1), 14–30.

Trumbull, D. J., Booney, R., & Grudens-Schuck, N. (2005). Developing materials to promote inquiry: lessons learned. *Science Education, 89*(4), 879–900.

Van den Akker, J., Gravemeijer, K., McKenney, S., & Nieveen, N. (2006). Introducing educational design research. In J. van den Akker, K. Gravemeijer, S. McKenney, & N. Nieveen (Eds.), *educational design research* (pp. 1–7). London: Routledge.

Van der Schee, J. A., & Rijborz, J. D. (2003). Coaching students in research skills: a difficult task for teachers. *European Journal of Teacher Education, 26*(2), 229–237.

Van der Valk, T., & De Jong, O. (2009). Scaffolding Science Teachers in Open-inquiry Teaching. *International Journal of Science Education, 31*(6), 829–850.

Van der Valk, A. E., & Van Soest, M. F. N. (2004). *Onderzoek leren doen in de bètavakken. Elementen van een leerlijn in de onderbouw van twee scholen. [Learning to do inquiry in science. Elements of a learning trajectory in the first classes of two schools]*. Utrecht: Universiteit Utrecht (CDβ/Onderwijskunde/ICO-ISOR).

Van Oers, B. (1998). From context to contextualizing. *Learning and Instruction, 8*(6), 473–488.

Van Oers, B. (2001). Contextualisation for abstraction. *Cognitive Science Quarterly, 1*(3–4), 279–306.

Van Rens, E. M. M. (2005). *Effective chemical education for learning to inquire in upper secondary schools.* Vrije Universiteit Amsterdam: PhD Thesis. http://www.naturfagsenteret.no/esera/phd/abstract51.html (accessed September 2009).

Van Rens, E. M. M., Pilot, A., & Van Dijk, H. (2004). Enhancement of quality in chemical inquiry by pre-university students. *International Journal of Science and Mathematics Education, 2*(4), 493–509.

Van Zee, E. H. (2000). Analysis of a student-generated inquiry discussion. *International Journal of Science Education, 22*(2), 115–142.

Vygotsky, L. S. (1978). Interaction between learning and development. In M. Cole, V. John-Steiner, S. Scribner, & E. Souberman (Eds.), *Mind in society. The development of higher psychological processes* (pp. 79–91). Cambridge: Harvard University Press. Original work published 1935.

Windschitl, M. (2003). Inquiry projects in science teacher education: what can investigative experiences reveal about teacher thinking and eventual classroom practice? *Science Education, 87*(1), 112–143.

Windschitl, M. (2004). Folk theories of "inquiry": how preservice teachers reproduce the discourse and practices of an atheoretical scientific method. *Journal of Research in Science Education, 41*(5), 481–512.

Witte, R. S., & Witte, J. S. (2004). *Statistics* (7th ed.). Hoboken: John Wiley & Sons, Inc.