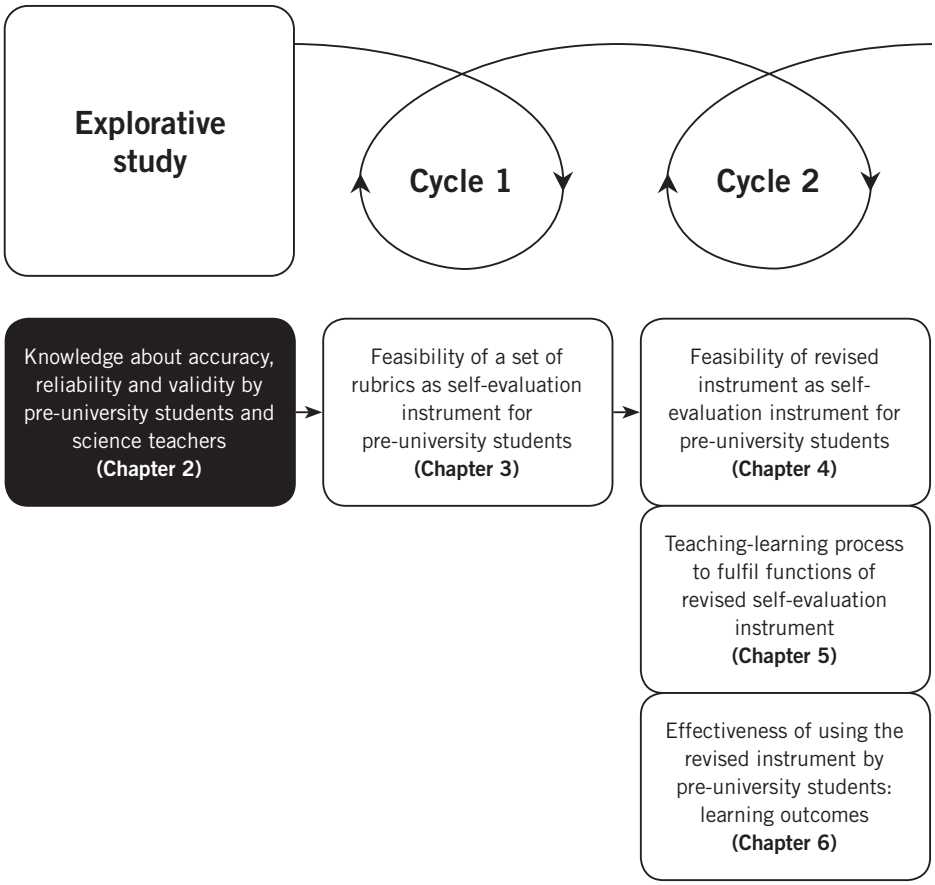

KNOWLEDGE ABOUT ACCURACY, RELIABILITY AND VALIDITY BY PRE-UNIVERSITY STUDENTS AND SCIENCE TEACHERS

CHAPTER 2

The objective of this study was to explore the extent to which pre-university science students and biology, physics and chemistry teachers recognise concepts of evidence (CoE) in an inquiry report of two students. These CoE help to enlarge the students' procedural understanding and their ability to ensure the accuracy, reliability and validity (ARV) of inquiries in biology, chemistry and physics at pre-university level. Pre-university science students and teachers of different science disciplines were involved in this qualitative study. Six students and six teachers first judged a student's inquiry report by completing a think-aloud task. Afterwards, the students were interviewed to survey their knowledge of ensuring ARV in scientific inquiries. In addition, 38 pre-university science students (aged 16 or 17) filled out a questionnaire on the meaning of ARV. The results showed that students recognise CoE that are necessary for improving the ARV of an inquiry to the same extent as teachers, but with a different focus. The findings of this study were used in the design of a self-evaluation instrument, a teaching-learning process with inquiry tasks to teach students to evaluate the ARV in inquiries in different school science subjects (Chapters 3–6).

This chapter is based on the article published as: Van der Jagt, S., Schalk, H., & Van Rens, L. (2011). Teachers' and students' use of concepts of evidence in judging the quality of an inquiry. In A. Yarden & G. S. Carvalho (Eds.), *Authenticity in biology education: Benefits and challenges. A selection of papers presented at the 8th Conference of European Researchers in Didactics of Biology (ERIDOB)* (pp. 41-52). Braga, Portugal.

2.1 THEORETICAL BACKGROUND

In many upper secondary school curricula, pre-university science students learn to perform inquiries in biology, chemistry and physics. When students perform an inquiry, they need to have at least some understanding of substantive facts, to be able to use necessary practical skills and to have some procedural understanding of how to construct a fair test and how to use evidence in reasoning (Figure 2.1) (Gott & Duggan, 2007; Schalk, Van der Schee, & Boersma, 2007). Nevertheless, in science curricula this procedural understanding is often trivialised. Most curricula associate procedural understanding only with goals such as communication of knowledge claims (Abd-El-Khalick et al., 2004).

In performing inquiries, students make observations or conduct experiments in order to find an answer to the inquiry question. When they draw a conclusion, they have to argue why the obtained results lead to that conclusion and not to another one. As Lawson (2010, p.2) points out, this line of reasoning is not primarily aimed at convincing others, but 'rather as one of discovering which of several possible explanations of a particular puzzling observation should be accepted and which should be rejected'.

Recent educational research in chemistry (Van Rens, Pilot, & Van der Schee, 2010) and biology (Schalk, Van der Schee, & Boersma, 2009) shows that the use of concepts of evidence (CoE) (Gott, Duggan, Roberts, & Hussain, n.d.) is convenient to improve students' procedural understanding. Some of the 82 described CoE are feasible to evaluate the validity, reliability and accuracy of an inquiry in the school science subjects (Gott & Duggan, 2003), but it is still unclear which of these CoE can be used by pre-university students.

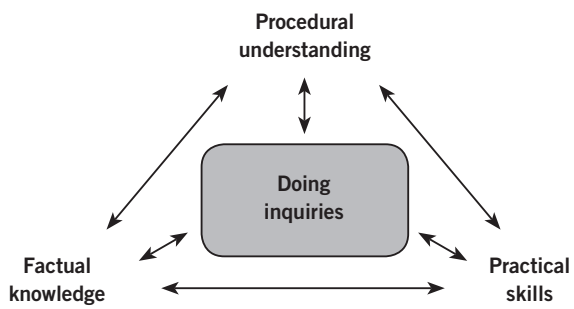

Figure 2.1
A model for performing inquiries in the school science subjects

It can be expected that when students at pre-university level perform inquiries in biology, chemistry and physics, their procedural understanding will improve when they learn how to use the CoE to ensure accuracy, reliability and validity (ARV) in all of these science subjects (Gott & Duggan, 2003). However, accuracy, reliability and validity have slightly different meanings in the different disciplines. In biology, for example, results can only be reliable when a representative sample of a population has been taken. To get reliable results in an experiment in, for example, some physics inquiries, one has to repeat the measurements. These different meanings can confuse students, especially when the analogy in supporting empirical reasoning in both inquiries is not made explicit to them (Millar, Driver, Leach, Scott, & Wood-Robinson, 1987).

For all inquiries with the objective of improving students' procedural understanding, the teacher has to choose which CoE are relevant and make them explicit to the students. To do this in an appropriate way, teachers should know how to interpret and use the CoE in each specific science inquiry in their science discipline. For this explorative study, we selected relevant CoE based on ensuring quality in biology, physics and chemistry inquiries. This selection is hereafter termed 'relevant CoE'.

2.2 KEY OBJECTIVES AND HYPOTHESIS

This explorative study is part of a larger study on learning to inquire in the school science disciplines. It aims to gain insight into how pre-university science students can increase their abilities to apply their procedural understanding on the evaluation of ARV in inquiries flexibly in different school science subjects. To reach this aim, an instrument with which pre-university science students can be supported in learning to evaluate the ARV of inquiries in different school science subjects had to be developed. This explorative study is focused on identifying CoE that can be used in designing this instrument and a teaching-learning process with inquiry tasks that will support students in evaluating the ARV of an inquiry in the different school science subjects. The following research questions guided this explorative study:

(1) *To what extent do pre-university science students and biology, physics and chemistry teachers recognise CoE that can be related to accuracy, reliability and validity in a student's inquiry?*
(2) *What is the accordance between biology, chemistry and physics teachers in the recognition of CoE in relation to accuracy, reliability and validity?*
(3) *What do pre-university science students know about the meaning of accuracy, reliability and validity in inquiries?*

The hypothesis was that biology, chemistry and physics teachers would make more use of relevant CoE to judge the ARV of a student inquiry than students would. According to Germann and Aram (1996), experts and novices differ in their knowledge and application of procedural understanding and practical skills. Therefore, it was

expected that all teachers would be more expert in judging the ARV of inquiries than students, due to their holding academic degrees in science disciplines and their previous experience with inquiries in the classroom.

2.3 METHOD

The knowledge of pre-university science students and science teachers about ARV of inquiries was determined by a qualitative study (Denscombe, 2007) that involved a think-aloud task (Bowen, 1994), interviews and a questionnaire.

2.3.1 Participants

Think-aloud task and interviews

Six upper pre-university science students (aged 16 or 17) from three different schools and six science teachers from five different schools in the Netherlands voluntarily participated in a think-aloud task. They received a bookshop gift card as incentive for participating in this study. The students and science teachers were from different schools. The teachers were asked by the researcher to participate in this study during a teachers' conference. All teachers held a master's degree, had at least five years of teaching experience at pre-university level and were familiar with students conducting inquiries and with judging students' inquiry reports. Two of them were biology teachers, two taught chemistry and two were physics teachers.

Three other teachers, from the network of the researcher, each asked two pre-university science students from their schools to participate in the think-aloud task and interviews. All participating students were studying biology, physics and chemistry at pre-university level. Five of the students were interviewed about their knowledge of the meaning of ARV in inquiries after performing the think-aloud task. One student was not interviewed because he had to get back to class directly after the think-aloud task.

Questionnaire

Thirty-eight pre-university science students (aged 16 or 17) from two different schools in the Netherlands filled out a questionnaire on the meanings of ARV in inquiries. All the students were familiar with conducting inquiries in school science subjects.

2.3.2 Materials and procedure

Think-aloud task

For the think-aloud task, a student-written ('scientific') article about the influence of chocolate on the physiology of the human body was selected. Two students (aged 16) had written this article as part of a chemistry inquiry project (Van Rens, Van der Schee, & Pilot, 2009). This student-written article can be found in Appendix A. Henceforth, the student authors of the article are termed 'student researchers'; 'students' refers to those students who participated in our explorative study.

Interviews and questionnaire

The interviews and the questionnaire consisted of six questions:

1) What does accuracy or inaccuracy mean, in your opinion?
2) What does reliability or unreliability mean, in your opinion?
3) What does validity or invalidity mean, in your opinion?
4) On which aspects should a researcher focus when evaluating the accuracy of an inquiry?
5) On which aspects should a researcher focus when evaluating the reliability of an inquiry?
6) On which aspects should a researcher focus when evaluating the validity of an inquiry?

These questions were formulated by two researchers and their validity was checked by three other researchers. Both the interview and the questionnaire were introduced by the researcher to elucidate the knowledge of the students about ARV of inquiries. Before filling out the questionnaire, the students were instructed by the researcher to write down all their thoughts and ideas about the ARV of inquiries.

2.3.3 Data collection

Think-aloud task

The participating students and teachers were asked to judge the quality of the inquiry as described in the student-written article by thinking aloud. The meaning of 'quality of an inquiry' was not specified. The participants were stimulated to speak out their thoughts and were asked to clarify their judgements. These questions were: 'Can you explain what you mean by… [repeating their wording]?' and 'About which part of the text did you just talk?' The responses of the participants were audio-recorded and transcribed verbatim.

Interviews and questionnaire

The responses of the interviews with students were audio-recorded and transcribed verbatim. The questionnaires were handed out by the researcher and collected directly after the students had answered the questions.

2.3.4 Data analysis

Think-aloud task

Before analysing the students' and teachers' judgements during the think-aloud task, the student-written article was analysed by two researchers to determine the relevant CoE that were, or should have been, used by the student researchers. These relevant CoE were selected from previous research by Gott et al. (n.d.) and Schalk (2006), and reformulated for the student-written article.

In this and the following studies, the rephrased descriptions of the relevant CoE in this specific student inquiry report are called 'items' in order to distinguish them from 'relevant CoE'. This list of items was independently constructed by two researchers, with an interrater agreement of 96%. For the rest of the items, a discussion took place until consensus was reached (Janesick, 2000). The list ultimately consisted of 47 items which are feasible for evaluating the accuracy, reliability or validity of an inquiry (see Appendix B).

Thereafter, the students' and teachers' responses from the think-aloud task were independently categorised as one of the 47 items by two researchers. The agreement in the categorisation of the two researchers was 84%. Differences in categorisation were discussed until consensus was reached. Subsequently, the mentioned items were related to the overarching concepts of accuracy, reliability and validity.

Interviews and questionnaire

The answers from the interviews and questionnaires about evaluating ARV when performing inquiries were independently categorised by two researchers as:

1) Use of correct and relevant CoE
2) Use of incorrect, but relevant CoE (e.g., while answering the question about accuracy a participant referred to a relevant CoE that contributes to reliability)
3) Use of everyday language to describe the concept being asked about (e.g., accuracy is determined as 'doing something neatly')
4) Nonsense/ambiguous answer
5) No answer

Sometimes an answer consisted of more than one statement; if necessary, these were categorised independently. The proportion agreement between the two researchers for the scores of the responses in the interviews was 75% and on the questionnaire 86%. Different scores were discussed until consensus was reached (Janesick, 2000).

2.4 RESULTS

2.4.1 Judging the quality of an inquiry

Table 2.1 contains an overview of the total number of items and of the number of items on validity, reliability and accuracy mentioned by the students and teachers during the think-aloud task as well as the averaged numbers and percentages for students and teachers. The number of items mentioned ranged between 10 and 16 different items per student. For the teachers this range was between 7 and 23 different items per teacher. The participants could mention at most 47 different items: 22 about validity, 13 about reliability and 12 about accuracy (see Appendix B for an overview of these 47 items).

Table 2.1
Overview of the total number of mentioned items by six students and six science teachers and of the averaged numbers and percentages

	Average of the six students	**Average of the six science teachers**
Total number of mentioned items (n=47)	13.5 (28.7%)	15 (31.9%)
Number of mentioned items about validity (n=22)	6.5 (29.5%)	7.3 (33.2%)
Number of mentioned items about reliability (n=13)	3.3 (25.4%)	3.5 (26.9%)
Number of mentioned items about accuracy (n=12)	3.7 (30.8%)	4.2 (35.0%)

When we look at these results qualitatively, some items were mentioned more clearly and more often by the participants than others. For example, five students mentioned logical reasoning in the theoretical framework of the judged student-written inquiry report, the measurement intervals of the independent variables and the importance of keeping other influential variables constant. It is remarkable that four of these five students repeated that other variables should be kept constant a couple of times during the think-aloud task. Four of the students mentioned the contribution of a good inquiry question to the validity of the inquiry and the importance of using the results of the inquiry when drawing a conclusion. About ensuring reliability, four students recognised that the student researchers did a control experiment before starting the first experiment on blood pressure (see Appendix A), but none of them stated that the student researchers neglected to do the same when conducting the second experiment. Four of the students mentioned that the student researchers should have repeated their measurements, but it was not clear whether all of them precisely understood the importance of repeating measurements to find the spread around an average measurement value.

Five teachers referred to the quality of the inquiry question and pointed to the evaluation of the accuracy of measurements of pupil dilation in the student-written article, although none of them stated that the student researchers had neglected to evaluate the accuracy of the blood pressure measurements. Four teachers judged the quality of the theoretical framework. They all mentioned the logical reasoning in the theoretical framework as well as the validity of the relation of this framework to the subsequent parts of the inquiry. Four teachers stated that the tables and graphs needed appropriate headings corresponding to the kind of information shown, and also expressed the importance of making recommendations for further inquiries. In addition, four teachers said that it is important to keep other influential variables constant, to justify the chosen sample size and explain its representativeness, and to repeat the measurements. Although three teachers referred to using the results from the inquiry as evidence for the conclusion, and restricting the conclusion to the

evidence from the inquiry, none of them talked about the validity of the conclusion relative to the inquiry question.

The students and teachers had a more or less equal score, but it was revealed that two-thirds of the total number of items mentioned by all students together were directly recognisable in the student-written article. The science teachers talked more often about items that were neglected by the student researchers.

2.4.2 Differences between the teachers

The differences and similarities between the teachers' results in light of their science discipline are summarised in Table 2.2. It should be kept in mind that this comparison is based on the results of three times two teachers and only gives an indication of the differences in focus of teachers from the different school science subjects.

Table 2.2
Comparison of the results of teachers from the different school science subjects

		Number of items about validity (n=22)	Number of items about reliability (n=13)	Number of items about accuracy (n=12)	Total number of items (n=47)
Chemistry	Teacher 1	5	4	2	11
	Teacher 3	8	4	6	18
Physics	Teacher 2	2	1	4	7
	Teacher 4	4	1	3	8
Biology	Teacher 5	12	7	4	23
	Teacher 6	13	4	6	23

The main difference between the biology, chemistry and physics teachers is seen in their judgements about items that help ensure the validity of the inquiry. The biology teachers' scores were highest in all categories, and they appeared to follow the empirical reasoning in the article more than the teachers from the other disciplines. Furthermore, the two chemistry teachers looked at the inquiry at different levels. Teacher 1 looked mainly at the practical skills of the student researchers while teacher 3 made comments on their procedural understanding. For example, referring to the graphs, teacher 1 said:

- *The results, they did it quite neatly, drawing a table, I always tell students: first draw a table and then a matching graph. I don't see many problems with it.*

Teacher 3 looked at the same graphs and focused on validity in terms of consistency between the data in the graphs and the conclusion. He said:

- *It is clear that you have to draw a graph to visualise the results. You can directly notice that white chocolate has no effect on the first human test subject, because the blood pressure fluctuates all the time. And the blood pressure of the other human test subject ended lower, after rising first. Actually, they didn't measure a lowering of the blood pressure as they state in the conclusion.*

2.4.3 Knowledge of the students

About accuracy

During the interviews, one student used a correct and relevant CoE to explain the meaning of accuracy in scientific inquiries: Read out the measurement values with different people. Two students also used everyday terms to answer the question about accuracy. All five interviewed students mentioned that repeatability and reproducibility of the inquiry can be related to the accuracy of an inquiry.

Similar results showed up in the questionnaire: four of the students used a correct and relevant CoE to explain how accurate measurement values can be obtained during inquiries. Three students wrote down a CoE that helps ensure reliability instead of accuracy:

- *It is important to keep other influencing variables constant … every time you repeat the experiment.*

Thirty-two students answered in everyday terminology in the questionnaire, such as:

- *Doing something neatly.*

About reliability

Two interviewed students made use of correct and relevant CoE to explain how the reliability of an inquiry can be ensured. One of them mentioned the influence of diversity in a sample population on the reliability of an inquiry; the other stated that a researcher has to be sure that the variety in outcomes is not caused by measurement errors. Three interviewed students referred to CoE that help ensure validity or accuracy. All students explained in everyday language that *reliability can be ensured when researchers use sources of information that can be trusted.*

In the questionnaire, four students wrote down a correct relevant CoE to explain how to ensure reliability. Eight students used a CoE that deals with the validity of an inquiry. Twenty-eight of the students used everyday language to describe how to ensure reliability, mostly by giving synonyms (in Dutch) for *trustworthy and faithful* or by writing down the isolated words 'facts' and 'theories' without specifying their importance to reliability.

About validity

During the interviews, four students referred to correct relevant CoE that help ensure the validity of an inquiry, for example, by saying:

- *In any case, the conclusion should be an answer to the inquiry question and it has to be based on the results of the inquiry you have done.*

One student mentioned a CoE about ensuring reliability:

- *You have to keep the other variables constant.*

One student used daily life terminology to describe how the validity of an inquiry can be maintained. These results differed from the answers in the questionnaire. Fifteen of the respondents in the questionnaire did not give an answer. None of the other students made use of a CoE to explain the meaning of validity. Sixteen students wrote down everyday language such as *performed without any shortcomings*. It sounds like these students were thinking of disabled people—in Dutch: 'invalide'.

In Table 2.3 an overview is given of the knowledge of students about accuracy, reliability and validity in inquiries as analysed in the interviews and questionnaires.

Table 2.3
Knowledge of students (n=43) about accuracy, reliability and validity of inquiries

Student's answer categorised as	Number of answers about accuracy	reliability	validity
Correct and relevant CoE	5	7	4
Incorrect, but relevant CoE	8	11	0
Everyday language	34	33	17
Nonsense/ambiguous answer	2	3	2
No answer	0	0	15

Note. The results of the analysis of the interviews and questionnaire are taken together. Some students gave more than one answer.

Knowledge versus performance

Most of the students interviewed demonstrated a lack of congruence between their knowledge about ensuring the quality of an inquiry and their performance during the think-aloud task. For example, during the interview, one student explained how to deal with more than two variables:

- *To get reliable results, you have to keep all variables constant, except the ones that are important for the inquiry.*

However, when judging the student-written article, she did not notice the use of more than two variables and its influence on the results and conclusion.

2.5 CONCLUSION AND INTERPRETATION

First, it can be concluded that students and teachers make use of relevant CoE in equal amounts to evaluate the ARV of an inquiry, but have focus differently on the recognisable versus neglected items in the student-written article. This result does not correspond to our expectation that teachers – because of their experience – would mention more relevant CoE than students. It is surprising that students know such a lot about ensuring the quality of an inquiry. Second, there is less accordance in the recognition of CoE by the teachers from the different school science subjects. Both biology teachers recognised far more CoE than the physics and chemistry teachers. Third, about the meaning of ARV in inquiries, it can be concluded that pre-university science students have some appropriate knowledge, but not enough to evaluate the ARV of an inquiry appropriately.

A possible explanation for this result can be found by looking at the role of inquiries in the science curricula in the Netherlands in the past few decades. Inquiry projects became a major part of the curriculum around 1996. Teachers were expected to enact inquiry teaching in class, but were not all trained on how to integrate these projects into their teaching. A minority of teachers completed a training course to become more expert in teaching students to inquire. As shown in the study of Gyllenpalm, Wickman and Holmgren (2009), teachers do not differentiate in their reasoning between methods of inquiry and methods of communication about inquiry. In our study, the same hybridisation can be recognised: while judging the quality of the inquiry, teachers often talk about the communication aspects of the products of inquiries. Based on this, it is not surprising that most teachers do not use CoE while judging the quality of a student-written article.

The difference between the teachers from different school science subjects can be interpreted by looking at inquiries in their disciplines in more detail. While performing an inquiry in biology at pre-university level, one frequently has to deal with the natural variation of (formerly) living test objects. A biological researcher always has to deal with limited reproducibility and repeatability of an inquiry because of this natural diversity (Pantin, 1968). It is possible that ensuring ARV in the present curricula at pre-university level is taught more explicitly in biological inquiries than in chemistry or physics. Consequently, the teachers might have judged the student-written article as they are used to judging the inquiries in their own subjects. This corresponds to a study of the UK Science Community Representing Education: most science teachers at pre-university level are confident in the inquiries of their own discipline, but are not convinced of their teaching skills in inquiries in other school science subjects. This problem appears to be more serious in physics than in biology or chemistry (SCORE, 2008).

About the meaning of ARV in inquiries, it can be concluded that pre-university science students have some appropriate knowledge. Looking at the results of the questionnaire and interviews, the interviewed students demonstrated better abilities to mention correct relevant CoE about ensuring validity than the 'questionnaire students'. The results on accuracy and reliability were equal. A methodological reason for the difference in the validity results can be that the interviewed students were more focused on scientific reasoning than the others because of the think-aloud task they had performed just before the interview. In retrospect, it would have been better if the students had been interviewed before starting the think-aloud task to minimise this influence. On the other hand, in the interviews, students mostly mentioned CoE other than those they had recognised in the student-written article, as illustrated in section 2.4.3: the student was probably told about influencing variables without improving her procedural understanding of the importance of controlling these variables in an inquiry.

In sum, these findings are not disappointing where students are concerned, but it appears that the knowledge of science teachers has to be expanded to give them the opportunity to teach how to ensure the quality of an inquiry more completely, with more reference to procedural knowledge and more coherently with the performance of students during inquiries.

2.6 LIMITATIONS

Because of the small number of participants, we have to be careful about generalising our conclusions. We tried to optimise the results by choosing participants from different schools. Another weakness of this study is that the conclusions are based on the performance during a think-aloud task and on interview results, so we do not know how students use evidence in the classroom while inquiring. However, the conclusions on judging the quality of an inquiry are similar to results of previous studies in the Netherlands and Sweden (Gyllenpalm et al., 2009; Schalk & Yuksel, 2009). Furthermore, although we selected a representative example of a student-written inquiry report from a database[2], it can be questioned whether the students and teachers would have arrived at similar results if another student-written inquiry report had been used in the think-aloud task. Finally, we realise that understanding CoE influences the understanding of the nature of scientific knowledge. However, we did not incorporate this broader perspective into our study, although it would have been interesting.

2 This selection was based on mistakes frequently made by pre-university students.

2.7 IMPLICATIONS

The results and conclusions of this explorative study were used to determine hypothetical design principles for the next research cycles about evaluating the ARV of inquiries in different school science subjects. More knowledge about students' and teachers' knowledge about CoE provided a starting point for the design of a self-evaluation instrument (Arter & McTighe, 2001; Germann & Aram, 1996) and an accompanying teaching-learning process with inquiry tasks.

Three researchers independently reflected on the 47 CoE items (see Appendix B) relative to the results with the criterion 'is the item appropriate for an elaboration in a student self-evaluation instrument regarding four phases in an inquiry'. In case of a disagreement discussion between the researchers took place until consensus was reached (Janesick, 2000). This reflection led to 23 CoE items that seem to be appropriate for self-evaluation of the ARV by pre-university students during the enactment of successive school science inquiry units. Table 2.4 gives an overview of the 23 CoE items.

Table 2.4
Overview of the 23 items from the CoE model that could be appropriate for pre-university students in a self-evaluation of ARV during the enactment of successive school science inquiries

Item(s) from the CoE model	
After preparing the inquiry plan	Logical reasoning in theoretical framework
	Specific and concrete inquiry question
	Hypothesis can be tested by inquiry method
	Conduct control experiment
	Sample is sufficiently large
	Sample is sufficiently varied
	Tables and graphs to sum up the results
	Measurement apparatus is sufficiently accurate
After collecting the data	The same independent variable throughout
	Reduce influence of other variables
	Measure or observe with more than one observer
	Measure or observe in an objective way
	Measure or observe in a systematic way
	Repeat measurements, calculate average and deviation
After handling the data	Measuring apparatus is calibrated
	Compare results with other inquiries
	Conclusion is based on inquiry results
	Inquiry question is answered in the conclusion
	Measuring instrument has an adequate range
After completing the inquiry	Inquiry method fits inquiry question
	Sufficient results to infer conclusion
	Conclusion fully fits the inquiry question and inquiry method
	Give recommendations for further inquiries

In addition, the teaching-learning process with the learning inquiry tasks that need to be designed needs to bridge the gap between the students' focus on recognisable and neglected items. For example, teachers first talk about the CoE items that the students recognise and then about how, if necessary, to improve the ARV of their inquiry. Next, they could give students feedback on items they neglected.

It remains important in a new situation to figure out participating students' prior knowledge about ensuring the ARV of an inquiry. This information can be used to decide on the next steps in improving students' understanding of ARV. Teachers need training to gain more insight into how to teach students the various CoE items relative to the ARV in an inquiry in order to ensure quality in the different school science inquiries.

REFERENCES

Abd-El-Khalick, F., BouJaoude, S., Duschl, R., Lederman, N. G., Mamlok-Naaman, R., Hofstein, A., et al. (2004). Inquiry in science education: International perspectives. Science Education, 88(3), 397–419.

Arter, J., & McTighe, J. (2001). Scoring rubrics in the classroom. Thousand Oaks, CA: Corwin Press.

Bowen, C. W. (1994). Think-aloud methods in chemistry education. Journal of Chemical Education, 71(3), 184–190.

Denscombe, M. (2007). The good research guide for small-scale social research projects (3rd ed.). Maidenhead, UK: Open University Press.

Germann, P. J., & Aram, R. J. (1996). Students' performances on the science processes of recording data, analyzing data, drawing conclusions, and providing evidence. Journal of Research in Science Teaching, 33(7), 773–798.

Gott, R., & Duggan, S. (2003). Understanding and using scientific evidence. London: SAGE Publications.

Gott, R., & Duggan, S. (2007). A framework for practical work in science and scientific literacy through argumentation. Research in Science & Technological Education, 25(3), 271–291.

Gott, R., Duggan, S., Roberts, R., & Hussain, A. (n.d.). Research into understanding scientific evidence. Retrieved on 5 November 2013 from http://www.dur.ac.uk/rosalyn.roberts/Evidence/cofev.htm

Gyllenpalm, J., Wickman, P., & Holmgren, S. (2009). Teachers' language on scientific inquiry: Methods of teaching or methods of inquiry? International Journal of Science Education, 32(9), 1151–1172.

Janesick, V. J. (2000). The choreography of qualitative research design. In H. K. Denzin & Y. S. Lincoln (Eds.), Handbook of qualitative research (pp. 379–399). Thousand Oaks, CA: Sage Publications.

Lawson, A. E. (2010). Basic inferences of scientific reasoning, argumentation, and discovery. Science Education 94(2), 336–364.

Millar, R., Driver, R., Leach, J., Scott, P., & Wood-Robinson, C. (1987). Beyond processes. Studies in Science Education, 14, 33–62.

Pantin, C. F. A. (1968). The relations between the sciences. London and New York: Cambridge University Press.

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. Retrieved from http://igitur-archive.library.uu.nl/dissertations/2006-1206-200836/index.htm

Schalk, H. H., Van der Schee, J. A., & Boersma, K. T. (2007). The development of understanding of evidence in pre-university biology education in the Netherlands. Paper presented at the 7th ESERA conference, August, Malmö, Sweden.

Schalk, H. H., Van der Schee, J. A., & Boersma, K. T. (2009). The use of concepts of evidence by students in biology investigations: Development research in pre-university education. In M. Hammann, K. Boersma and A. J. Waarlo (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. Utrecht: Beta Press.

Schalk, H. H., & Yuksel, A. (2009). The use of concepts of evidence in argumentations about the quality of investigations. Paper presented at the 8th ESERA conference, September, Istanbul, Turkey.

SCORE. (2008). Practical work in science. A report and proposal for a strategic framework. London: Science Community Representing Education.

Van Rens, L., Pilot, A., & Van der Schee, J. (2010). A framework for teaching scientific inquiry in upper secondary school chemistry. Journal of Research in Science Teaching, 47(7), 788–807.

Van Rens, L., Van der Schee, J., & Pilot, A. (2009). Teaching molecular diffusion using an inquiry approach. Diffusion activities in a secondary school inquiry-learning community. Journal of Chemical Education, 86(12), 1437–1441.