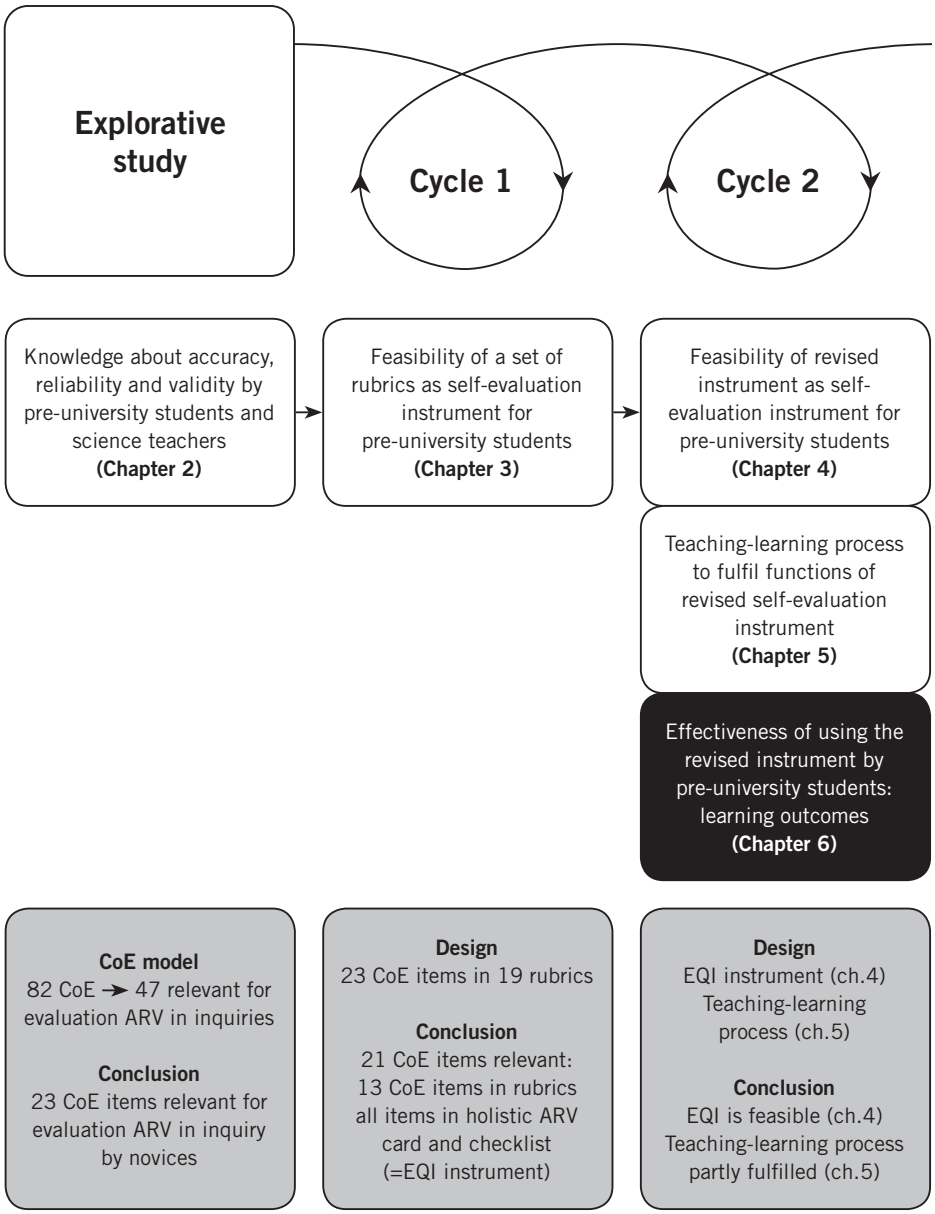

EFFECTIVENESS OF USING THE REVISED INSTRUMENT BY PRE-UNIVERSITY STUDENTS: LEARNING OUTCOMES

CHAPTER 6

This chapter presents a study on the student learning outcomes (effectiveness) of the previously described teaching-learning process (see Chapter 5) in which students used rubrics to evaluate accuracy, reliability and validity (ARV) in inquiries in different school science subjects. These rubrics are part of the Evaluating the Quality of Inquiries (EQI) instrument. To determine the effectiveness of using this instrument, four learning aims were set: relevance of the instrument for evaluation purposes, knowledge about the content, application in a new inquiry, and transfer from one inquiry to another. Twenty-seven pre-university science students used the instrument in four successive inquiries. In the first three inquiries, students worked in groups on planning, performing and evaluating inquiries about the cooling of coffee (general science unit), human taste (biology unit) and speed and braking distance of cyclists in a traffic situation on a dangerous crossing (physics unit). In the last unit they worked individually on the evaluation of ARV in an inquiry about chocolate and on the planning of an inquiry about tooth decay. Questionnaires, interviews and written documents were used to obtain data from students and the teacher. It was concluded that students perceive the EQI instrument as relevant for evaluating the ARV in inquiries, gain sufficient knowledge about concepts of evidence, and, with the support of peers and teacher, apply these concepts in a new inquiry. Four inquiry assignments seemed to be insufficient to reach transfer of using the EQI instrument to evaluate the ARV of inquiries without guidance of peers and teacher. Further research is needed on the effect of the EQI instrument regarding transfer.

This chapter is based on the article: S. A. W. Van der Jagt, L. Van Rens, H. H. Schalk, A. Pilot and J. J. Beishuizen. *Learning to evaluate the quality of inquiries at pre-university level by using a self-evaluation instrument.* In preparation.

6.1 INTRODUCTION

In line with a worldwide trend, the development of inquiry skills is included in the Dutch formal pre-university science curriculum. This formal curriculum involves the development of student insight into the nature of science and the way scientists work and think (Abd-El-Khalick et al., 2004; Van Rens, Van Muijlwijk, Beishuizen, & Van der Schee, 2013). Learning to inquire has various aims. This study focuses on procedural understanding about evaluating accuracy, reliability and validity (ARV) in inquiries (Gott & Duggan, 1995; Millar, 2010). Increasing their procedural understanding helps pre-university science students to get an idea of the criteria and practices that scientists use when they plan, conduct and evaluate their research (Chinn & Malhotra, 2002). Pre-university science students are novices in this domain. They can hardly give meaning to ARV let alone evaluate ARV in an inquiry in school science subjects (Lunetta, Hofstein, & Clough, 2007; Schalk, Van der Schee, & Boersma, 2013; and see Chapter 2). Moreover, these students seldom recognise that there are many similarities in planning and conducting inquiries, as well as evaluating ARV, in different inquiries (Roberts & Gott, 2002).

Transfer of concepts can increase when students evaluate their inquiries in a structured way and when differences in the use of concepts in various inquiries are made explicit for students (Bransford, 2000; Gilbert, Bulte, & Pilot, 2011). To increase their ability to transfer, it also seems necessary to let students formulate their own inquiry questions and let them design their own inquiries (O'Neill & Polman, 2004). Salomon and Perkins (1989) distinguished two ways to improve transfer: a lot of practice *(low-road transfer)* and mindful abstraction *(high-road transfer)*. Bransford (2000) concluded that novices, such as pre-university science students when evaluating ARV in an inquiry, should perform learning activities in which they can recognise domain-specific patterns and are guided to link the new knowledge to their prior knowledge *(high-road transfer)*.

Andrade and Valtcheva (2009) showed that the use of a self-evaluation instrument 'obliges' students to reflect on and revise their knowledge and abilities in different learning domains. Self-evaluation can be useful in focusing the attention of the students on the most important aspects of a task, by helping the students to structure knowledge or by showing them the strong and weak parts of their abilities. A self-evaluation instrument can also be useful in the way that students gain knowledge in one school science subject and apply it flexibly to another subject.

Sevian and Gonsalves (2008) suggested that a coherent set of rubrics could function as a self-evaluation instrument. Rubrics support learning by making performance criteria explicit. These criteria make it easier for peers or the teacher to evaluate, assess and give feedback when students perform self-evaluation of their work with rubrics (Jonsson & Svingby, 2007).

However, the review study of Tierney and Simon (2004) showed that rubrics that have been developed previously for evaluating student inquiries do not lead to satisfactory improvements in novices' evaluation of ARV in inquiries. Such rubrics contained vague descriptions like: 'You found sufficient literature references that fit the inquiry question'. From such a description it is unclear for the students how many references are considered 'sufficient', or when references 'fit' the inquiry question. For novices, such descriptions are too vague for them to determine the quality of their own performances let alone think of improvements. Moreover, Jonsson and Svingby (2007) concluded in their review study that most rubrics focus on the assessment of the quality of students' essays and reports and not on evaluation of their inquiry processes.

This calls for the development of a process-oriented and clearly structured set of rubrics, with which pre-university science students can evaluate ARV in inquiries. The set of rubrics developed in this study is called the Evaluation of Quality of Inquiries (EQI) instrument. For the full description of the EQI instrument, see Appendix C. In the second research cycle, empirical support for the feasibility of this instrument was found (see Chapter 4) and 15 hypothetical design characteristics (hDCs) were identified for a teaching-learning process in which the EQI instrument can be used for evaluation of ARV in inquiries of different school science subjects (see Chapter 5). In this chapter a third study, as part of the second research cycle, is described: the research on the effectiveness of the use of the EQI instrument in the teaching-learning process, or the influence on student learning outcomes. In this teaching-learning process, the EQI instrument was used to determine whether students could apply knowledge about evaluating ARV in inquiries flexibly in different school science subjects.

6.2 THEORETICAL FRAMEWORK

To determine the content of the EQI instrument, the concepts of evidence (CoE) model, as described by Gott, Duggan, Roberts and Hussain (n.d.), appeared to be suitable. This model proved its strength in the development of learning and teaching activities that improved student procedural insight related to the concepts of accuracy, reliability and validity in inquiries in the science subjects (Schalk, Van der Schee, & Boersma, 2009; Van Rens, Pilot, & Van der Schee, 2010).

Gott et al. (n.d.) described 82 concepts of evidence (CoE) that are important in the process of collecting empirical evidence in a scientific inquiry. As described in Chapter 2, 47 CoE are related to the evaluation of ARV of inquiries. These CoE were rewritten for use by pre-university students and hereafter called CoE items. From the explorative study (see Chapter 2) it was concluded that 23 of these 47 could be elaborated in a self-evaluation instrument for novices. In the first research cycle (see Chapter 3), 21 CoE were found feasible for pre-university science students in evaluating ARV in inquiries. Of these 21 CoE items, 13 were worked out in twelve rubrics. These

rubrics appeared to support pre-university students, novices, to a sufficient extent in evaluating ARV in an inquiry. During the first research cycle, it was observed that novices easily lost their way when using rubrics and that they needed an overview of the CoE items. That is why an ARV card (holistic overview tool) showing the connection between all used CoE in a completed inquiry became part of the EQI instrument. Moreover, an easy-to-use checklist was incorporated in the instrument. Ultimately, the rubric Theoretical framework was not used in the studies in the second research cycle because in the successive inquiry units the theoretical framework was given to the students. Consequently, eleven rubrics – and as a consequence their accompanying 12 CoE items – were included in the successive inquiry units. Table 6.1 presents an overview of all 20 items from the CoE model in the EQI instrument used in this study. Six items concern CoE items of accuracy, six of reliability and eight items are assigned to the CoE items of validity.

Jonsson and Svingby (2007) concluded that with a hierarchical description in rubrics, students and teachers can assess performance levels in an inquiry and can decide on what to do to improve their performance. The hierarchical descriptions in rubrics can be based on the Structure of Observed Learning Outcomes (SOLO) taxonomy (Biggs & Tang, 2007). This taxonomy supported undergraduate science students when they evaluated their performances at various points during a learning task (e.g. Hodges & Harvey, 2003; Levins & Pegg, 1993; Minogue & Jones, 2009).

The SOLO taxonomy distinguishes five performance levels of increasing complexity: prestructural, unistructural, multistructural, relational and extended abstract. A prestructural level occurs when students use daily language, for example, to describe accuracy: 'we did everything as precisely as possible'. When students use one aspect, mostly by imitation of the teacher's language, then the level of performance is considered unistructural. On the multistructural level of performance, the students describe relevant aspects, but do not consider inconsistencies or possible relations between these aspects. On the relational level of performance, students do consider inconsistencies or possible relations. When students are able to indicate how inconsistencies and related aspects fit in the inquiry domain concerned, they perform on the extended abstract level. If the SOLO taxonomy is correctly elaborated in an instrument it is expected that a certain level of performance can only be achieved by students when they have mastered all the previous levels of performance of the SOLO taxonomy (Biggs & Tang, 2007). Regarding the issue of how detailed the EQI instrument needs to be, most studies show that a detailed and analytical instrument is the most suitable for self-evaluation by students who are novices in a certain domain (e.g. Arter & McTighe, 2001; Mertler, 2001).

Table 6.1
Overview of the 20 items of the CoE model as elaborated in the EQI instrument and used in the successive inquiry units

Items from CoE model used to evaluate ARV in an inquiry	
Accuracy	1. Measure or observe with more than one observer
	2. Measure or observe in an objective way
	3. Measure or observe in a systematic way
	4. Measurement apparatus is sufficiently accurate
	5. Measuring instrument has an adequate range
	6. Measuring apparatus is calibrated
Reliability	1. Reduce influence of other variables
	2. Conduct control experiment
	3. Repeat measurements, calculate average and deviation
	4. Sample is sufficiently large
	5. Sample is sufficiently varied
	6. Compare results with other inquiries
Validity	1. The same (in)dependent variable throughout
	2. Specific and concrete inquiry question
	3. Hypothesis can be tested by inquiry method
	4. Inquiry method fits inquiry question
	5. Sufficient results to infer conclusion
	6. Conclusion is based on inquiry results
	7. Inquiry question is answered in conclusion
	8. Give recommendations for further inquiries

The EQI instrument should be functional for students in evaluating ARV in various inquiries, so the instrument should also be generic for use in all the intended domains. Crucial in this is that the effects of transfer become visible when the inquiries are, in the eyes of the students, sufficiently related (Beishuizen & Asscher, 2001). Jonsson and Svingby (2007) found that students understand generic descriptions better when these are accompanied by norm-referenced benchmark samples. In a previous study (see Chapter 3) it was concluded that these examples should be from one theme that covers the three school science subjects, general enough for the students to be able to make a transfer to various inquiries, and complex enough to be specified at all SOLO taxonomy levels.

Furthermore, according to Gagné, Yekovich and Yekovich (1993), transfer can be stimulated when students are explicitly asked to write down and discuss issues with peers. At the same time, the aim should be to let students develop their conceptual knowledge, because this facilitates transfer. Whether a student can indeed

transfer achieved knowledge and abilities on the evaluation of ARV to a new inquiry depends on his or her ability and motivation to transfer knowledge and skills (Broad & Newstrom, 1992; Van Oers, 1998).

The three functions of the EQI instrument in the student learning process can be comprehended as:

(1) support student self-evaluation;
(2) support evaluation by peers and teacher; and
(3) support transfer from one to another inquiry.

Earlier in the second test cycle of this study, it was shown that a design based on these three functions can lead to the intended teaching-learning process regarding the first and second functions (see Chapter 5). The experiment on the testing of the use of the EQI instrument in the teaching-learning process was also used for gathering data for the present study, although some aspects of the teaching-learning process need minor revisions (see Table 5.6, p. 135).

Hence, the research question of the present study was: *What is the effectiveness of the EQI instrument in the transfer of ARV evaluation skills in various inquiries performed by pre-university science students?*

To determine the effectiveness of the EQI instrument, four learning aims were set by the researchers, based on the above described previous research studies:

- Learning aim 1 *Relevance:* Students perceive relevance when they learn how to evaluate the ARV in an inquiry and perceive that the EQI instrument is helpful to do so.
- Learning aim 2 *Knowledge:* Students learn enough items of the CoE model and are able to evaluate ARV in an inquiry adequately.
- Learning aim 3 *Application:* Students use, with support from teacher and peers, items of the CoE model in a new inquiry at the same or a higher level in the SOLO taxonomy.
- Learning aim 4 *Transfer:* Students transfer, without guidance from teacher or peers, items of the CoE model to a new inquiry at a sufficient level in the SOLO taxonomy.

Therefore, the research question contained the following sub questions:

1) *To what extent do students perceive that the EQI instrument is relevant for evaluating ARV in an inquiry?*
2) *To what extent do students have sufficient knowledge to evaluate ARV in an inquiry?*
3) *To what extent can students apply this achieved knowledge on how to evaluate ARV in a, for them, new inquiry at the same or a higher level in the SOLO taxonomy?*

4) *To what extent can students, without any guidance, transfer achieved knowledge on the evaluation of ARV to a, for them, new inquiry at a sufficient level in the SOLO taxonomy?*

6.3 METHOD

The effectiveness of the use of the EQI instrument was determined by a mixed methods research approach (Denscombe, 2007) with triangulation of data (Yin, 2003). Observations and written materials were analysed to determine the student learning outcomes regarding the evaluation of ARV in successive inquiry units.

6.3.1 Participants

Twenty-seven pre-university science students, aged 16 or 17, voluntarily participated in the second test cycle (see also Chapter 4 and 5). They signed up to participate after receiving an e-mail from their science teachers about this study. All students were studying biology, physics and chemistry at pre-university level at the same school. They all had experience with practical work in the science subjects, but they had never systematically evaluated ARV in an inquiry. The students received a small financial incentive after finishing the last inquiry session to reward them for their effort in participating in the study after regular school periods. The teacher is not a unit of analysis in this study. More information about the role of the teacher can be found below (see 6.3.2).

6.3.2 Materials and procedure

The students worked in twelve groups of two or three students on three successive inquiry units: general science, biology and physics (the same units as used in the studies as described in Chapters 4 and 5). The composition of the groups was the same in all units. The students worked on the three units in seven afternoon sessions, each of three hours, over a time span of three months. The decision to depart from regular lesson periods of 50 minutes was made because the students' inquiries required more time. Moreover, the interruptions that often happen in regular school periods were avoided.

Each group worked on the inquiry units successively, formulating inquiry plans with inquiry questions, hypotheses and inquiry methods related to the specified problems, which were: the cooling rate of hot coffee (general science unit); taste on the human tongue (biology unit); and improving a traffic situation on a dangerous crossing by measuring the speed and braking distance of cyclists (physics unit). In each inquiry unit, each group evaluated the ARV of their inquiry plans, with the use of the EQI instrument. The teacher also evaluated the students' inquiry plans with the instrument and noted down feedback. With this feedback at hand, the students discussed possible improvements and wrote their final inquiry plans. After a check on safety by

the teacher, the students carried out their plans and worked the results and empirical evidence out into conclusions. After this, the groups used the instrument again to evaluate ARV in the entire inquiry.

During the last afternoon, the students worked individually, without any guidance from the teacher or their peers, for one hour on an assessment regarding a chemistry inquiry. In this assessment, the students completed two paper-and-pencil tasks. With the instrument at hand, the students evaluated ARV in the report of a student inquiry on the effect of eating various types of chocolate on pupil dilation and blood pressure (the chocolate inquiry). In the second task, the tooth decay inquiry, students first formulated an inquiry question related to a theoretical exposé of the effect of beverages with a low pH-value on dental enamel. After handing in their answers, they received the inquiry question, 'To what extent is dental enamel better protected by saliva that arises during chewing than by saliva that keeps the mouth wet?'. The students were supposed to write down a hypothesis and an inquiry method for how to answer this question. The chocolate and tooth decay inquiries have proved to be sufficiently open and motivating for pre-university chemistry students (Van Rens et al., 2010) and were therefore suitable for this study. The two assessment tasks, one on the evaluation of ARV in a student inquiry on the effect of chocolate on blood pressure and pupil dilation and the other one on writing an inquiry plan to investigate the effect of acid in drinks on tooth enamel or tooth decay, are available on http://hdl.handle.net/1871/38422. Figure 6.1 shows an overview of the successive inquiries and the two assessment tasks.

Beforehand, it should be noticed that both the inquiry tasks and the assessment tasks could have a high level of complexity for students because of the mixture of learning

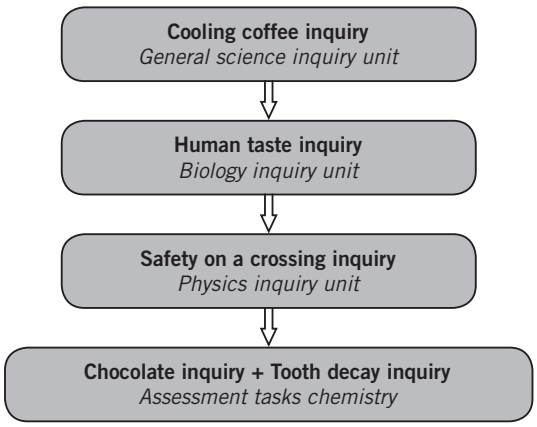

Figure 6.1
Overview of the successive inquiries in the various inquiry units and assessment task

to evaluate the ARV and at the same time the application of this knowledge to new inquiries. A comparison of the levels of difficulty of the successive units and tasks was beyond the scope of this explorative study on the use of the EQI instrument for transfer purposes, but it should be kept in mind as influencing the learning outcomes (this aspect is further elaborated in the discussion, see 6.5)

The researcher – who holds a master's degree in biology and a degree in teaching, and has eight years of experience in pre-university biology teaching – was one of the designers and the teacher of the various inquiry units in the second test cycle. Being one of the designers avoided misinterpretation by the teacher of both the content of the instrument and the intended teaching-learning process. Moreover, having one teacher to teach all the inquiry units reduced the distortion of continuity that would result from three different teachers being involved in the teaching-learning process. In the research procedure some precautions were taken to prevent the 'dual role' of teacher and researcher becoming a limitation of the study. The first precaution was the involvement of two independent observers in all the afternoon sessions, who observed whether the lessons were taught as planned and whether the teacher gave any answers that could directly influence the learning outcomes. The second precaution was to analyse the data independently by two researchers.

The role of the teacher was to introduce the instrument preceding the first use of it by the students, to explain the terminology used in the instrument, and to instruct, guide and support the students all the time they were using the instrument in the three successive inquiry units. For each of the inquiry units a workbook with activities for the students was designed. All inquiry units were also guided with a teacher's manual in which the intended use of the student EQI as well as the guides for instruction to enact the unit were described (see http://hdl.handle.net/1871/38422).

6.3.3 Data collection

The data collection will be described for each of the four learning aims.

Learning aim 1: Students perceive relevance when they learn how to evaluate the ARV in an inquiry and experience that the EQI is helpful to do so.

A self-report method was used to determine whether the first learning aim had been achieved. The students individually completed a questionnaire with three questions immediately after they finished the cooling coffee inquiry and after the tooth decay inquiry assessment task. These were two closed questions with a five-point Likert scale: (1) indicate the relevance of learning how to evaluate ARV, and (2) indicate the relevance of the EQI instrument for doing so. Moreover, a third and open question was posed: What do you think of working with the EQI instrument?

Furthermore, one week after the last afternoon session all twelve student groups were interviewed by two researchers who had not been present in the seven learning sessions. One of the questions asked was: What do you think of the relevance of the EQI instrument in evaluating ARV in the various inquiry units? All the interviews were audio-recorded and transcribed.

Learning aim 2: Students learn enough items of the CoE model and are able to evaluate ARV in an inquiry adequately.

To determine how many items the students had learned, all the students' evaluations with the EQI instrument of the chocolate inquiry from the assessment task and of their cooling coffee inquiries in the first inquiry unit were collected.

Learning aim 3: Students use, with support from teacher and peers, items of the CoE model in a new inquiry at the same or a higher level in the SOLO taxonomy.

The students' inquiry questions and hypotheses and the methods of their inquiries concerning the cooling coffee, human taste, safety on a crossing and the tooth decay inquiries were collected to determine whether the third learning aim was reached. The researchers filled out the rubrics from the EQI instrument to determine at which level of complexity the students used the items of the CoE model in the respective inquiries.

Learning aim 4: Students transfer, without guidance from teacher or peers, items of the CoE model to a new inquiry at a sufficient level in the SOLO taxonomy.

The students' levels of performance, as scored by the researchers (see previous paragraph) in the inquiry questions, hypotheses and inquiry methods in the tooth decay, safety on a crossing, human taste and cooling coffee inquiries were used to determine whether the fourth learning aim, transfer of knowledge on items of the CoE model to a new inquiry, was reached.

6.3.4 Data analysis

To determine the student learning outcomes each individual student was taken as a unit of analysis, although they had worked in groups during most of the learning sessions (Kock, Taconis, Bolhuis, & Gravemeijer, 2013). All data analysis was conducted independently by two researchers and in all cases the proportion agreement was good (≥77%). When differences occurred a discussion took place until consensus was reached (Janesick, 2000).

The data analysis for each of the four learning aims is described below.

Learning aim 1: Relevance

For the first learning aim, the percentage of student responses – scores 4 and 5 on the Likert scale – after the cooling coffee and tooth decay inquiries were determined

for both questions 1 and 2. Moreover, all student responses to the open question 3 of the questionnaire were read and weighted as to whether they could be categorised as 'student thinks the EQI instrument is relevant' or as 'student thinks the EQI instrument is irrelevant'. The transcripts of the group interviews were read and the responses were categorised as positive or negative regarding the relevance of the EQI instrument for evaluating ARV in an inquiry. The proportion agreement between both researchers in the scoring of responses was 77%.

To determine whether the first learning aim was achieved by the students, it was decided that at the time that the students had completed the last inquiry unit at least 80% (≥ 22 students) should have indicated that the EQI instrument is relevant in evaluating the ARV in an inquiry (cf. Juran, Gryna, & Bingham, 1974). The latter point was chosen because at that time the students had used the instrument frequently, and in different inquiries, and would probably give a balanced response on the relevance of the instrument.

Learning aim 2: Knowledge

For the second learning aim, all student responses in the evaluation of the ARV in the chocolate inquiry, as well as their responses in the evaluation of the ARV in their cooling coffee inquiries, were categorised and determined (number, %) by:

1. Item of EQI regarding accuracy adequately used
2. Item of EQI regarding reliability adequately used
3. Item of EQI regarding validity adequately used
4. Item of EQI used in a different way than in the CoE model
5. Item of the CoE model used that was not included in the EQI instrument
6. Use of everyday language regarding ARV
7. Unclear response

The proportion agreement between both researchers in categorising the responses was 89%. When more than 22 students scored at least 80% (cf. Juran et al., 1974) of the maximum sum score over categories 1, 2 and 3 (i.e., at least 16 out of 20), then students had learned enough items from the CoE model to be able to evaluate ARV in an inquiry adequately.

In a further analysis, the responses on ARV of each student in the chocolate inquiry as well as in the cooling coffee inquiry in each of categories 1, 2 and 3 was determined. For an achieved learning outcome (see Table 6.1) in category 1, regarding accuracy, as well as in category 2, regarding reliability, students should in both cases show at least five out of the six CoE items (≥80%) in their responses. Regarding validity, category 3, at least seven out of eight CoE items (≥80%) should occur in their responses (cf. Juran et al., 1974).

Learning aim 3: Application

For the third learning aim, each student's inquiry question, hypothesis and method in the tooth decay, safety on a crossing, human taste and cooling coffee inquiries was independently scored by two researchers with the EQI instrument. They had a proportion agreement of 81%. This was done to determine whether the students used items from the CoE model adequately in a new inquiry with the same or a higher SOLO taxonomy level of complexity. The scores 1–5 corresponded with the five levels – prestructural (level 1) to extended abstract (level 5) – in the SOLO taxonomy. The number of students with a higher or equal level of performance in the inquiry question, hypothesis and inquiry method of each of the three inquiries – tooth decay, safety on a crossing and human taste – compared with the cooling coffee inquiry was first determined. The norm for achievement of learning aim 3 was set, again in line with Juran et al. (1974), at 22 students or more (≥ 80%) who should at least perform better when comparing the levels of performance in the inquiry question, hypothesis and inquiry method in the four inquiries.

Learning aim 4: Transfer

To determine whether learning aim 4 was achieved, first the researchers compared the students' levels of performance, according to the SOLO taxonomy, in the inquiry question, hypothesis and inquiry method (from learning aim 3) in the tooth decay inquiry, safety on a crossing inquiry, human taste inquiry and cooling coffee inquiry. Then the number and percentage of students who performed on the relational level 4 or on the extended abstract level 5 in the inquiry question, hypothesis and inquiry method in these four inquiries were counted. Levels 4 and 5 in the SOLO taxonomy are levels of performance in which all relevant aspects are related to each other (level 4) and are related to other relevant domains (level 5). The norm for achieving learning aim 4 was set at 22 students or more (≥ 80%; cf. Juran et al., 1974) who should at least perform on SOLO taxonomy level 4 regarding the inquiry question, hypothesis and inquiry method in the successive inquiries.

6.4 FINDINGS

6.4.1 Learning outcomes regarding aim 1: Relevance

In the questionnaire after the cooling coffee inquiry, ten students (37%) responded with a score of 4 or 5 on the Likert scale for the relevance of learning how to evaluate ARV in an inquiry and eight students (30%) for the relevance of the EQI instrument to do so. After the tooth decay inquiry, 26 students (96%) scored 4 or 5 for the first as well as the second question and one student response showed a score of 3 on both questions.

Analysis of student responses after the first inquiry unit on the third (open) question of the questionnaire showed that four students (15%) gave responses that can be categorised as relevant for evaluating ARV with the EQI instrument, whereas the re-

sponses of three students (11%) indicated irrelevance. The responses of nine students could not be categorised and eleven students did not respond to the question.

After the tooth decay inquiry, the analysis of the responses on the third question showed that the responses of 23 students (86%) can be categorised as involving 'relevance'. An example is:

- … *it* [the EQI instrument] *listed all points of how to go about an inquiry and what to pay attention to.*

Two students' responses could be categorised as 'irrelevant', for example:

- *It became boring to complete every time that whole bunch of papers, it was not meaningful to me anymore.*

Two students (7%) did not respond to question 3.

Analysis of the transcript of the interviews revealed that all students indicated the relevance of the EQI instrument while evaluating the ARV in an inquiry. An example is:

- *To conduct a good inquiry is quite difficult. At school you are not used to looking at the accuracy and so on … and then those papers [the EQI instrument] are very handy to see what you can still improve.*

In summary, more than 80% of the students perceived the relevance of the use of the EQI instrument in evaluating the ARV in an inquiry.

6.4.2 Learning outcomes regarding aim 2: Knowledge

Analysis of all student evaluative responses on the chocolate and cooling coffee inquiries with the EQI instrument respectively revealed 411 and 143 responses that could be categorised as adequate. The students' evaluative responses in the cooling coffee inquiry showed that 16 students (59%) made 18 adequate responses concerning the accuracy of the inquiry (category 1). In the chocolate inquiry, 26 students (96%) made 92 adequate responses. Moreover, the evaluative responses concerning accuracy in the chocolate inquiry showed that all six items from the CoE model related to the accuracy of an inquiry are mentioned by the students. Analysis of the evaluative responses regarding accuracy in the cooling coffee inquiry showed four of the six items. Twenty-four evaluative responses on accuracy in the chocolate inquiry concerned the CoE model item 'Measuring instrument is accurate enough' and 19 responses were related to both the items 'Measure and observe with more than one observer' and 'Measure and observe in a systematic way'.

Furthermore, 25 students (93%) made 104 adequate responses on reliability (category 2) in the chocolate inquiry and 20 students (74%) made 27 adequate responses in the cooling coffee inquiry. The evaluative responses concerning the reliability of the chocolate inquiry showed that all six items from the CoE model related to the reliability of an inquiry are mentioned by the students. Three items appeared in the evaluative student responses regarding reliability in the cooling coffee inquiry. There

were 23 evaluative responses on reliability in the chocolate inquiry concerned the item 'Sample is large enough'; 18 of these concerned both the items 'Reduce influence of other variables' and 'Sample is sufficiently varied'.

Moreover, 25 students (93%) made 130 adequate responses on validity (category 3) in the chocolate inquiry and 7 students (26%) made eight adequate responses in the cooling coffee inquiry. The evaluative responses showed that in the chocolate inquiry all seven items of the CoE model related to the validity of an inquiry are mentioned by the students. Analysis of the evaluative responses regarding the validity in the cooling coffee inquiry showed four of the seven items. On the validity of the chocolate inquiry, 23 adequate evaluative responses concerned the item 'Hypothesis can be tested by the inquiry method' and 20 the item 'Specific and concrete inquiry question'.

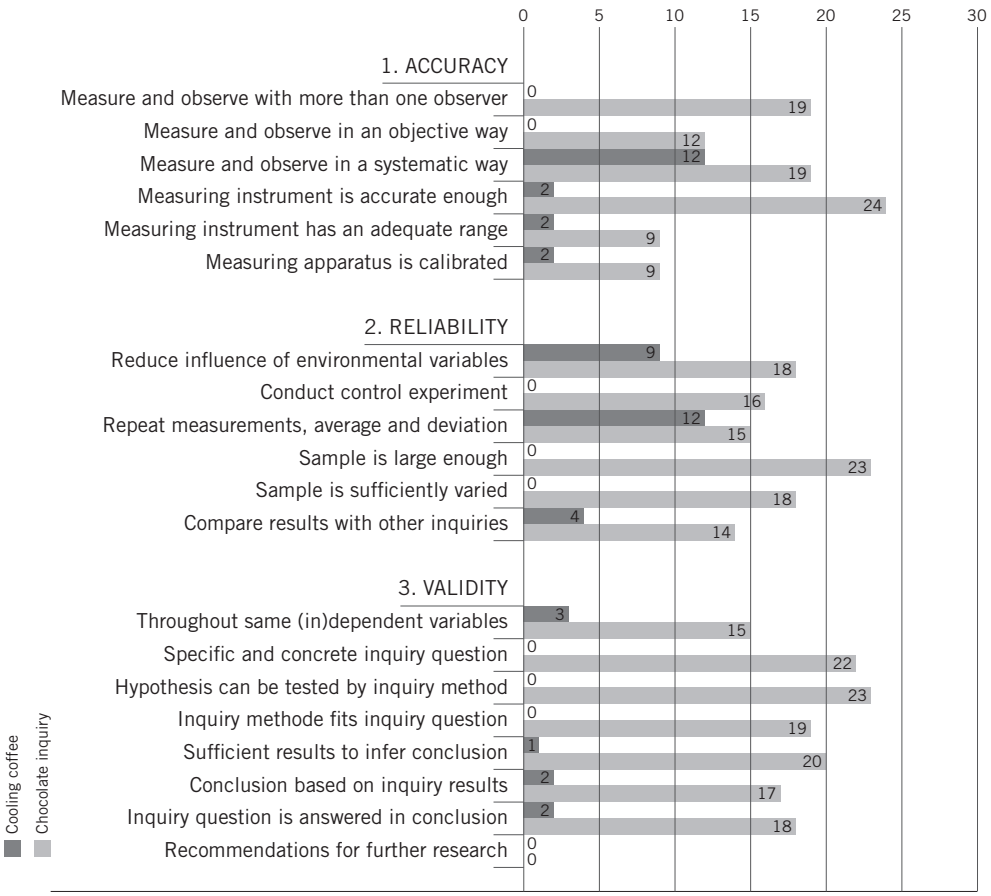

Figure 6.2

Overview of the number of students (n=27) giving adequate responses with the EQI instrument on the cooling coffee inquiry as well as on the chocolate inquiry regarding each of the CoE model items in the instrument that are related to accuracy, reliability and validity

In summary, in the chocolate inquiry more than 80% of the students adequately mentioned one or more items from the CoE model in each of the three categories on adequate evaluation of ARV. Figure 6.2 shows an overview of the number of students' evaluative responses with the EQI instrument on the chocolate inquiry as well as the cooling coffee inquiry regarding each of the 20 items in the instrument.

Twenty-three responses from 12 students (44%) and 34 responses from 23 students (85%) were categorised as items from the EQI instrument that are used differently than in the CoE model (category 4) in the analysis of the student evaluative responses on the chocolate and cooling coffee inquiries. For example, one student mentioned that *'The sample of two persons was too small'* in evaluating the validity of the chocolate inquiry. In the CoE model, this was seen as a part of the reliability of an inquiry. Moreover, the analysis of the students' evaluative responses on the two inquiries showed that eight responses made by six students (22%) and two responses made by two students (7%), respectively, are an item from the CoE model that does not appear in the EQI instrument (category 5), for example, *'Measurement values were conveniently arranged in graphs.'*

Regarding category 6, everyday language, analysis revealed 31 responses made by 15 students (56%) on the chocolate inquiry and 39 responses made by 23 students (85%) on the cooling coffee inquiry. Some examples are: *'You need to do it precisely'* (accuracy); and *'The inquiry should be conducted in an honest way'* (reliability). For category 7, unclear response, analysis revealed 23 responses made by 15 students (56%) on the chocolate inquiry and 15 responses made by 12 students (44%) on the cooling coffee inquiry. Some examples are: *'Clean materials';* and *'Graphs = mistake'*. An overview of the students' evaluative responses with the EQI instrument on the chocolate as well as the cooling coffee inquiry both in number and percentage in each of the categories 1–7 is presented in Table 6.2.

Table 6.2
Overview of the students' responses (n=27) in categories 1–7 (number and percentage) in the evaluation of ARV in two inquiries with the EQI instrument

Category	**Chocolate Inquiry**		**Cooling Coffee Inquiry**	
	Responses (%)	Students (n=27)	Responses (%)	Students (n=27)
1. Item EQI accuracy	92 (22%)	26	18 (13%)	16
2. Item EQI reliability	104 (25%)	25	27 (19%)	20
3. Item EQI validity	130 (32%)	25	8 (6%)	7
4. Item other than CoE	23 (6%)	12	34 (24%)	23
5. Item not in EQI	8 (2%)	6	2 (1%)	2
6. Daily language	31 (7%)	15	39 (27%)	23
7. Unclear response	23 (6%)	15	15 (10%)	12
Total	411 (100%)		143 (100%)	

Further analysis of the number of adequate evaluative responses on accuracy (category 1), reliability (category 2) and validity (category 3) in the chocolate inquiry as well as in the cooling coffee inquiry per student is presented in Table 6.3. It shows that the evaluative responses in the EQI instrument on the chocolate inquiry of 7 students (26%) contained five items concerning accuracy, those of 13 students (48%) contained five or more items concerning reliability, and those of 6 students (22%) contained seven or more items concerning the validity in an inquiry.

In summary, five students (19%) in the chocolate inquiry and none of the students (0%) in the cooling coffee inquiry reach the set norms of at least 80% of the items regarding accuracy, reliability and validity in the evaluative responses on the ARV.

Table 6.3

Overview of the number of evaluative responses of the students (n=27) in categories 1, 2 and 3 and the total number on the ARV in the chocolate inquiry (CH) and the cooling coffee (CC) inquiry

Student	Category 1 Accuracy (n=6)		Category 2 Reliability (n=6)		Category 3 Validity (n=8)		Total items (n=20)	
	CH	CC	CH	CC	CH	CC	CH	CC
1	3	0	5	1	6	1	14	2
2	3	1	2	2	3	0	8	3
3	2	1	5	2	6	0	13	3
4	3	0	3	1	7	1	13	2
5	4	1	6	2	6	1	16	4
6	2	0	5	0	3	0	10	0
7	4	0	4	0	7	0	15	0
8	4	1	6	0	3	0	13	1
9	4	0	5	1	6	2	15	3
10	5	0	6	1	5	0	16	1
11	0	1	0	1	0	0	0	2
12	2	0	3	0	3	0	8	0
13	1	1	1	1	1	0	3	2
14	5	1	2	0	5	1	12	2
15	5	0	5	1	4	0	14	1
16	3	0	4	2	6	0	13	2
17	4	1	4	1	5	0	13	2
18	4	1	5	0	6	1	15	2
19	1	0	0	1	0	0	1	1
20	5	3	6	2	7	0	18	5
21	4	1	2	1	6	0	12	2
22	5	1	5	2	7	0	17	3
23	4	1	2	2	5	0	11	3
24	5	0	4	1	7	0	16	1
25	5	1	5	1	4	0	14	1
26	3	0	5	0	7	0	15	0
27	2	1	4	1	5	1	11	3

Note. The shaded cells show the achieved norms of learning outcomes.

6.4.3 Learning outcomes regarding aim 3: Application

For the third learning outcome, data from the tooth decay inquiry, the safety on a crossing inquiry and the human taste inquiry were compared with the cooling coffee inquiry. Twenty-two students (81%) had higher or equal SOLO taxonomy levels of performance on the inquiry question and hypothesis compared with the previous inquiry unit. In the comparison of the hypothesis from the tooth decay versus the cooling coffee inquiry, 14 students showed an equal or higher level performance. Furthermore, the analysis showed that between 12 (44%) and 15 (56%) students had higher or equal SOLO taxonomy levels of performance on the inquiry method in the comparisons between the successive inquiries. For an overview of the comparisons, see Table 6.4 and Table 6.5.

Table 6.4
Number of students (n=27) with a higher or equal SOLO taxonomy level of performance in the four inquiries each compared with the cooling coffee inquiry

Comparison between the inquiries	**Students (n=27) with higher or equal performance level**	
	Number	**Percentage**
Inquiry question		
Tooth decay (4) – Cooling coffee (1)	21	78
Safety on a crossing (3) – Cooling coffee (1)	23	85
Human taste (2) – Cooling coffee (1)	23	85
Hypothesis		
Tooth decay (4) – Cooling coffee (1)	14	52
Safety on a crossing (3) – Cooling coffee (1)	22	81
Human taste (2) – Cooling coffee (1)	27	100
Inquiry method		
Tooth decay (4) – Cooling coffee (1)	15	56
Safety on a crossing (3) – Cooling coffee (1)	15	56
Human taste (2) – Cooling coffee (1)	12	44

In the tooth decay inquiry, 21 students (78%) had a higher or equal level in their inquiry question compared with their level of performance in the cooling coffee inquiry. An analysis of the level of performance on the hypothesis and the inquiry method analysis revealed that 14 (52%) and 15 (56%) students had a higher or equal level of performance respectively in the tooth decay inquiry, when compared with the cooling coffee inquiry.

In summary, the set norm of 80% of the students was achieved in all comparisons of the quality of the inquiry question, including the assessment task on tooth decay. The

set norm was also achieved in the comparisons of the quality of the hypothesis of the first and second, and the first and third, inquiry units.

Table 6.5

Comparison of the SOLO taxonomy levels (1–5) of performance of each student (n=27) in inquiry question, hypothesis and inquiry method in the tooth decay (TD) and cooling coffee (CC) inquiries

Student	Level of Inquiry Question		Level of Hypothesis		Level of Inquiry Method	
(n=27)	TD	CC	TD	CC	TD	CC
1	5	5	1	4	4	5
2	5	4	5	4	2	3
3	5	4	1	4	4	3
4	5	5	1	4	3	3
5	5	5	4	4	2	3
6	5	5	5	4	3	3
7	5	5	5	4	3	3
8	5	4	5	3	3	5
9	5	4	1	3	5	5
10	5	4	5	3	3	5
11	2	4	1	1	2	3
12	5	4	4	1	1	3
13	5	2	4	4	4	3
14	2	4	1	4	3	4
15	5	1	5	3	1	1
16	1	2	5	4	3	3
17	1	4	1	4	1	4
18	3	3	5	3	5	1
19	4	3	2	3	3	1
20	5	4	1	4	4	3
21	5	1	4	3	3	1
22	5	4	2	4	5	3
23	5	4	5	4	5	3
24	2	4	1	4	3	4
25	1	4	2	4	4	4
26	5	4	1	4	5	4
27	5	5	5	4	3	5

Note. The shaded cells show higher or equal level of performance in the TD inquiry unit compared with the level of performance in the previous CC inquiry unit.

6.4.4 Learning outcomes regarding aim 4: Transfer

Figure 6.3 shows the student SOLO taxonomy levels of performance in the tooth decay, safety on a crossing, human taste and cooling coffee inquiries regarding the inquiry question, hypothesis and inquiry method.

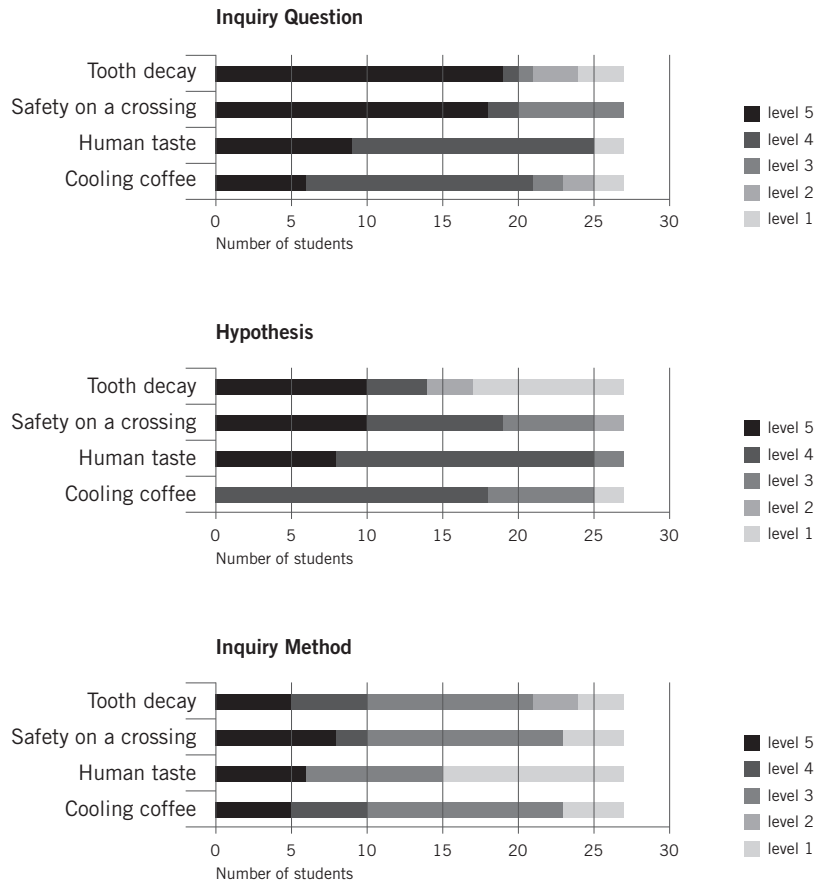

Figure 6.3
Student (n=27) SOLO taxonomy levels (1–5) of performance in four inquiries regarding the inquiry question, hypothesis and inquiry method

Regarding the inquiry question, 20 students (74%) in both the tooth decay and the safety on a crossing inquiries performed at SOLO taxonomy level 4 or 5 with averaged level scores of 4.1 and 4.4, respectively. In the human taste inquiry, 25 students (93%) showed level 4 or 5 with an average score over all students of 4.1, whereas 21 students (78%) showed these performance levels in the cooling coffee inquiry with an average score of 3.8.

Regarding the hypothesis, 14 (52%), 19 (70%), 25 (93%) and 18 students (67%) showed performance levels 4 or 5 of the respective successive inquiries with averaged level scores between 3.0 and 4.2. Concerning the inquiry method, 10 (37%), 10 (37%), 6 (22%)

and 10 students (37%) showed performance levels 4 or 5 of the respective successive inquiries with averaged level scores between 2.6 and 3.4. For an overview of these results, see Table 6.6. In summary, the set norm of 80% (22 students) was not reached in most inquiries.

Table 6.6
Number of students (n=27) with SOLO taxonomy levels of highest performance 4 or 5, averaged levels and standard deviation (s.d.) regarding the inquiry question, hypothesis and inquiry method in the four inquiries

Inquiries (last to first)	**Number of students (n=27) with performance level 4 or 5**		**Averaged level (s.d.) over all students (n=27)**
	Number	**Percentage**	
Inquiry question			
Tooth decay	20	74	4.1 (1.5)
Safety on a crossing	20	74	4.4 (0.9)
Human taste	25	93	4.1 (1.0)
Cooling coffee	21	78	3.8 (1.1)
Hypothesis			
Tooth decay	14	52	3.0 (1.8)
Safety on a crossing	19	70	4.0 (0.9)
Human taste	25	93	4.2 (0.6)
Cooling coffee	18	67	3.5 (0.8)
Inquiry method			
Tooth decay	10	37	3.2 (1.2)
Safety on a crossing	10	37	3.4 (1.3)
Human taste	6	22	2.6 (1.6)
Cooling coffee	10	37	3.3 (1.2)

6.5 DISCUSSION, CONCLUSION AND IMPLICATIONS

This study focused on students' learning outcomes when they use the EQI instrument to evaluate the ARV in successive inquiry units. The learning outcomes concern four aims: relevance, knowledge, application and transfer. The findings of this study showed that most students (more than the norm of 80%) perceived the relevance of using the EQI instrument in evaluating the ARV in an inquiry (learning aim 1). Also, most students (more than the norm of 80%) gained enough knowledge on items that are helpful in evaluating the ARV in an inquiry during the successive inquiry units (learning aim 2). Furthermore, most students (more than the norm of 80%) were able to apply knowledge about the ARV in an inquiry question in all inquiry units and about the ARV of the hypotheses in two inquiry units on a similar or higher level of

performance than in the first inquiry unit (learning aim 3). Finally, the set norm for transfer of evaluating ARV in inquiries at sufficient SOLO taxonomy levels of performance was not reached in this study. The findings are discussed below.

Regarding *learning aim 1 Relevance,* students think that it is relevant that they learn how to evaluate the ARV in an inquiry and think that the EQI instrument is helpful in doing so; it is concluded that this aim is achieved, because after the chocolate inquiry most of the students approved of both. The same applies for the student responses in the group interviews. However, some students indicate in the questionnaire that the relevance of the instrument decreases in the successive inquiries. These students may have gained a more cyclic way of thinking, as is common in scientific research, and as a result they develop a dislike for the structured approach of the EQI instrument or do not need to use the instrument anymore.

With regard to *learning aim 2 Knowledge,* students learn enough items about the CoE model to be able to evaluate ARV in an inquiry adequately and they know which items are suitable for evaluating ARV in an inquiry. Analysis of the evaluative student responses with the EQI instrument on the chocolate inquiry shows that most students (more than the norm of 80%) adequately evaluate the accuracy, reliability and validity of that inquiry (see Table 6.2: categories 1, 2 and 3). The students also use all CoE items from the EQI instrument (see Figure 6.2). Moreover, comparison of the evaluative student responses in the chocolate with the cooling coffee inquiry shows a large gain in student knowledge regarding ARV (see Table 6.2). The lowest number of adequate evaluative student responses concern the items 'Measuring instrument has an adequate range' and 'Measuring apparatus is calibrated' in category 1 on accuracy. Neither of these items was introduced explicitly to the students during the cooling coffee inquiry unit and, therefore, these items were probably not understood well enough by the students to be applied in the evaluation of their new inquiries. Novices need explicit examples from experts before they can apply new information in a new situation (Bransford, 2000; Oh, 2010). Hence the design of the modules requires some adaptations.

It is unknown to what extent the three inquiry units and assessment tasks have the same levels of difficulty for the students, because this could have influenced their results when applying the EQI instrument in the successive inquiries. Such a comparison on the equality in levels of difficulty is not easy to perform and is beyond the scope of this study. Despite the lack of such a comparison it can be tentatively concluded, from the detailed analysis of the evaluative responses in the chocolate inquiry versus the cooling coffee inquiry of each student, that the students gained most knowledge regarding reliability and validity in an inquiry (see Tables 6.2 and 6.3).

Regarding *learning aim 3 Application,* with support from teacher and peers, students used items of the CoE model in a new inquiry at the same or a higher level of the SOLO taxonomy. So, this aim is achieved by some of the students. The best student learning outcomes are visible in the inquiries on human taste and on the safety on a crossing versus the cooling coffee inquiry and regarding both the inquiry question and the hypothesis. This does not occur in the comparison of the students' SOLO taxonomy level of performance in the tooth decay inquiry versus the cooling coffee inquiry regarding the inquiry question (Table 6.4). An explanation could be that, in the tooth decay inquiry, the guidance of the teacher and peers is absent, whereas in the other inquiries the students gave each other feedback. Probably this guidance is still a necessity at this stage (Van Rens, Pilot, & Van Dijk, 2004). It is also possible that the content and context in the tooth decay inquiry, related to the effect of acidic solutions on dental enamel, is more complex and therefore more difficult for the students. Table 6.4 also shows that students made the least progress in formulating an inquiry method. As proved by Kuhn, Amsel and O'Loughlin (1988), the quality of a student's inquiry method depends on his or her expectations of the possible outcomes of the experiment. If students think of relatively simple outcomes, then the design of their inquiry method will be simple as well. This might be more complex and difficult in the subject matter of the final inquiries than in the first inquiries. Another reason for the little progress in the quality of the inquiry method in the successive inquiries could be that formulating an inquiry question and a hypothesis is less complex and contains fewer aspects than planning an inquiry method (Tamir, Stavy, & Ratner, 1998). Successful finishing of an inquiry plan also depends upon students' familiarity with the inquiry problem, the subject matter in the inquiry, their previous knowledge and experience with inquiries, and their communication skills. Writing an inquiry plan requires the students to be able to communicate their ideas completely and clearly (Germann, Aram, & Burke, 1996).

In addition, the content of the EQI instrument can be too narrowly aligned to the CoE so that using it in a complex inquiry situation has little impact on the individual understanding of the students (Hickey, Taasoobshirazi, & Cross, 2012). Therefore, it appeared that *with* the guidance of teacher and peers most of the students can apply items of the CoE model in another, and for them new, inquiry. Some students probably know the CoE items, but are unable to apply them (without support) in a new complex task. This is in line with other studies, for example, Hoekstra and Korthagen (2011), who concluded that a change from knowing to doing needs support from the teacher or peers.

For *learning aim 4 Transfer,* students transfer – without guidance from the teacher or peers – items from the CoE model in a new inquiry with a sufficient SOLO taxonomy level of complexity; it is concluded that this aim is not achieved. Many students

achieved equal or higher SOLO taxonomy levels of performance regarding the inquiry question and the hypothesis in the tooth decay inquiry compared with the cooling coffee inquiry (Table 6.6). However, the conduct and evaluation with the EQI instrument of three inquiries seem to be to insufficient for some students to achieve, without guidance, the relational or extended abstract level in evaluating ARV in a new inquiry (Bransford, 2000). The norm of 80% was not reached in the three goals for learning outcomes (Figure 6.3 and Table 6.6). Internalisation of criteria and the benchmark samples, and the development of self-evaluation skills is a complex process. Probably the students first need to get more expertise before the process of transfer can take place. More practice and feedback from the teacher or peers could probably help to achieve this (Andrade & Du, 2007; Pea, 1993). For this learning aim the same remark has to be made on the comparability of the level of difficulty of the different inquiries as is done in the previous discussion about learning aim 3.

From the above it can be concluded that the EQI instrument is an effective instrument with which the students learn and apply CoE items when they use the instrument to evaluate ARV in inquiries. Further study is needed to determine whether achievement of transfer can be brought about by more practice for the students with the EQI instrument in various inquiries with mindful abstraction, as suggested by Salomon and Perkins (1989). The same applies for the role of teacher and peers, which seems to be crucial in the effectiveness of the EQI instrument. More research is needed to find out how teacher and peers can indeed contribute to higher levels of transfer. Comparability of the difficulty of the tasks, including comparability regarding the inquiry question, hypothesis and inquiry method seems an important issue for further research and design.

REFERENCES

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

Andrade, H.,& Du, Y. (2007). Student responses to criteria-referenced self-assessment. Assessment and Evaluation in Higher Education, 32(2), 159–181.

Andrade, H., & Valtcheva, A. (2009). Promoting learning and achievement through self-assessment. Theory into Practice, 48(1), 12–19.

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

Beishuizen, J. J., & Asscher, J. J. (2001). Syllabus onderwijspsychologie. [Educational psychology workbook] (4th ed.). Leiden, The Netherlands: Department of Psychology, University of Leiden.

Biggs, J., & Tang, C. (Eds.). (2007). Teaching for quality learning at university (3rd ed.). Buckingham, UK: Open University Press.

Bransford, J. D. (2000). How people learn: Brain, mind, experience, and school (Expanded ed.). Washington, DC: National Academy Press.

Broad, M. L., & Newstrom, J. W. (1992). Transfer of training: Action-packed strategies to ensure high payoff from training investments. Reading, MA: Addison-Wesley.

Chinn, C. A., & Malhotra, B. A. (2002). Epistemologically authentic inquiry in schools: A theoretical framework for evaluating inquiry tasks. Science Education, 86(2), 175–218.

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

Gagné, E. D., Yekovich, C. W., & Yekovich, F. R. (1993). The cognitive psychology of school learning (2nd ed.). New York: HarperCollins.

Germann, P. J., Aram, R., & Burke, G. (1996). Identifying patterns and relationships among the responses of seventh-grade students to the science process skills of designing experiments. Journal of Research in Science Teaching, 33(1), 79–99.

Gilbert, J. K., Bulte, A. M. W., & Pilot, A. (2011). Concept development and transfer in context-based science education. International Journal of Science Education, 33(6), 817–837.

Gott, R., & Duggan, S. (1995). Investigative work in the science curriculum. Buckingham, UK and Philadelphia, PA: Open University Press.

Gott, R., Duggan, S., Roberts, R., & Hussain, A. (n.d.). Research into understanding scientific evidence Retrieved 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.

Hickey, D.T., Taasoobshirazi, G. & Cross, D. (2012). Assessment as learning: Enhancing discourse, understanding, and achievement in innovative science curricula. Journal of Research in Science Teaching, 49(10), 1240–1270.

Hodges, L. C., & Harvey, L. C. (2003). Evaluation of student learning in organic chemistry using the SOLO taxonomy. Journal of Chemistry Education, 80, 785–787.

Hoekstra, A., & Korthagen, F. (2011). Teacher learning in a context of educational change : Informal learning versus systematically supported learning. Journal of Teacher Education, 62(1), 79–93.

Hofstein, A., Navon, O., Kipnis, M., & Mamlok-Naaman, R. (2005). Developing students' ability to ask more and better questions resulting from inquiry-type chemistry laboratories. Journal of Research in Science Teaching, 42(7), 791–806.

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.

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. Educational Research Review, 2(2), 130–144.

Juran, J. M., Gryna, F. M., & Bingham, R. S. (Eds.). (1974). Quality control handbook (3rd ed.). New York: McGraw-Hill.

Kock, Z. J., Taconis, R., Bolhuis, S., & Gravemeijer, K. (2013). Some key issues in creating inquiry-based instructional practices that aim at the understanding of simple electric circuits. Research in Science Education, 43(2), 579–597.

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

Levins, L., & Pegg, J. (1993). Students' understanding of concepts related to plant growth. Research in Science Education, 23, 165–173.

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. 393–442). Mahwah, NJ: Lawrence Erlbaum Associates.

Mertler, C. A. (2001). Designing scoring rubrics for your classroom. Practical Assessment, Research & Evaluation, 7(25). Retrieved from http://PAREonline.net/getvn.asp?v=7&n=25

Millar, R. (2010). Analysing practical science activities to assess and improve their effectiveness. Hatfield, UK: The Association for Science Education.

Minogue, J., & Jones, G. (2009). Measuring the impact of haptic feedback using the SOLO taxonomy. International Journal of Science Education, 31(10), 1359–1378.

Oh, P. S. (2010). How can teachers help students formulate scientific hypotheses? Some strategies found in abductive inquiry activities of earth science. International Journal of Science Education, 32(4), 541–560.

O'Neill, D. K., & Polman, J. L. (2004). Why educate 'little scientists'? Examining the potential of practice-based scientific literacy. Journal of Research in Science Teaching, 41(3), 234-266.

Pea, R. D. (1993). Learning scientific concepts through material and social activities: Conversational analysis. Educational Psychologist, 28(3), 265–278.

Roberts, R., & Gott, R. (2002). Investigations: Collecting and using evidence. In D. Sang & V. Wood-Robinson (Eds.), Teaching secondary scientific enquiry. London: Association for Science Education.

Salomon, G., & Perkins, D. N. (1989). Rocky roads to transfer: Rethinking mechanisms of a neglected phenomenon. Educational Psychologist, 24(2), 113–142.

Schalk, H. H., Van der Schee, J., & 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 & 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 7th Conference of European Researchers in Didactics of Biology (ERIDOB). Zeist, The Netherlands. Utrecht: Bèta Press.

Schalk, H.H., Van der Schee, J., & Boersma, K. Th. (2013). The development of understanding of evidence in pre-university biology education in the Netherlands. Research in Science Education, 43(2), 551–578.

Sevian, H., & Gonsalves, L. (2008). Analysing how scientists explain their research: A rubric for measuring the effectiveness of scientific explanations. International Journal of Science Education, 30(11), 1441–1467.

Tamir, P., Stavy, R., & Ratner, N. (1998). Teaching science by inquiry: Assessment and learning. Educational Research, 33(1), 27–32.

Tierney, R., & Simon, M. (2004). What's still wrong with rubrics: Focusing on the consistency of performance criteria across scale levels. Practical Assessment, Research & Evaluation, 9(2). Retrieved from http://PAREonline.net/getvn.asp?v=9&n=2

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

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), 788–806.

Van Rens, L., 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 Rens, L., Van Muijlwijk, J., Beishuizen, J., & Van der Schee, J. (2013). Upper secondary chemistry students in a pharmacochemistry research community. International Journal of Science Education, 35(6), 1012–1036.

Yin, R. K. (2003). Case study research: Design and methods (3rd ed.). Thousand Oaks, CA: Sage Publications.