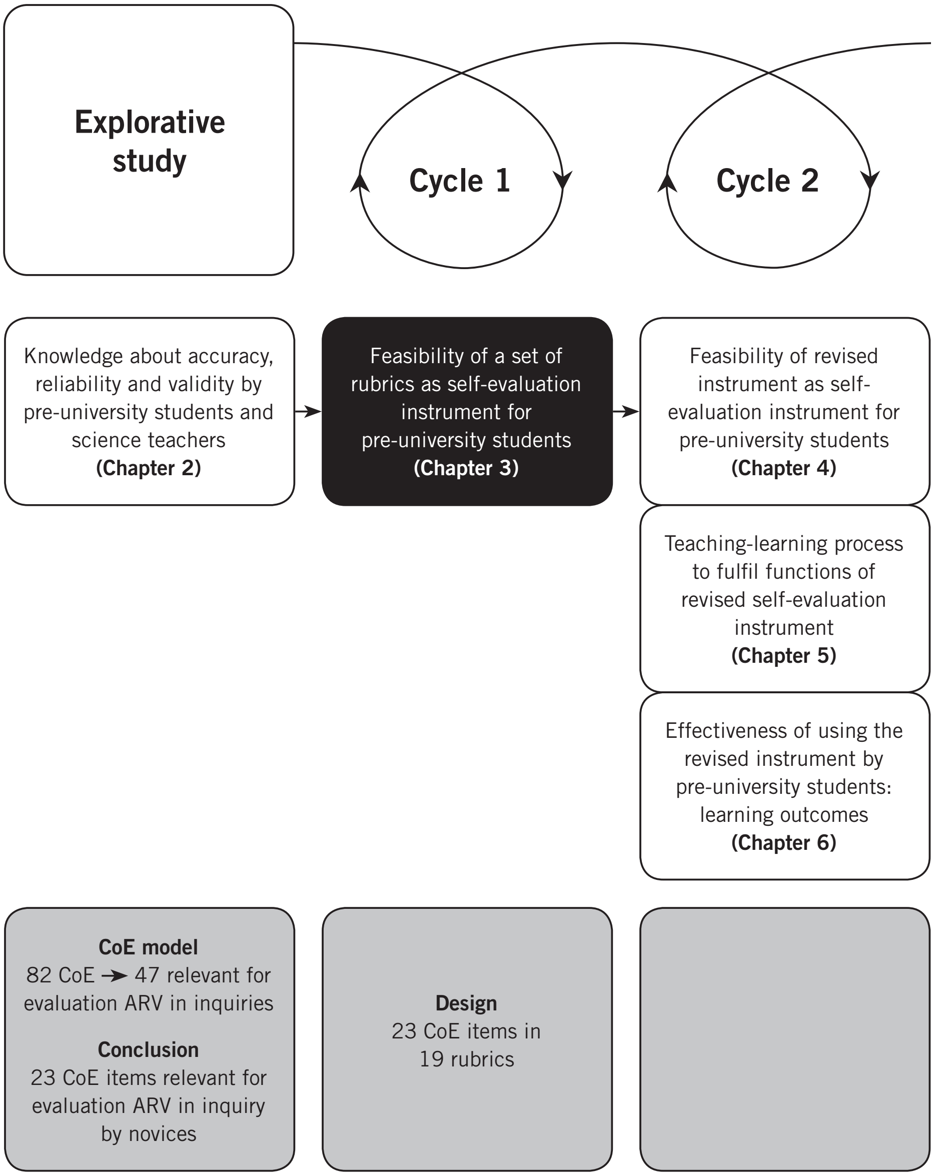

Explorative study

Cycle 1

Cycle 2

Knowledge about accuracy, reliability and validity by pre-university students and science teachers **(Chapter 2)**

Feasibility of a set of rubrics as self-evaluation instrument for pre-university students **(Chapter 3)**

Feasibility of revised instrument as self-evaluation instrument for pre-university students **(Chapter 4)**

Teaching-learning process to fulfil functions of revised self-evaluation instrument **(Chapter 5)**

Effectiveness of using the revised instrument by pre-university students: learning outcomes **(Chapter 6)**

CoE model
82 CoE → 47 relevant for evaluation ARV in inquiries

Conclusion
23 CoE items relevant for evaluation ARV in inquiry by novices

Design
23 CoE items in 19 rubrics

FEASIBILITY OF A SET OF RUBRICS AS SELF-EVALUATION INSTRUMENT FOR PRE-UNIVERSITY STUDENTS

CHAPTER 3

This study aimed at identifying design characteristics for a set of rubrics that is feasible for pre-university science students in learning how to evaluate the accuracy, reliability and validity (ARV) of inquiries that they conduct. Four hypothetical design characteristics: the content, the extent of complexity, the extent of detail and the extent of general application, were identified from the literature. These characteristics were used to design a set of 19 rubrics, based on 23 items from the concepts of evidence (CoE) model (see Chapter 2) and the Structure of Observed Learning Outcomes (SOLO) taxonomy. To determine the feasibility of the designed set, 24 pre-university students and two teachers from one school enacted the 19 rubrics in class in three successive science inquiry units. Data were obtained from written documents, questionnaires and interviews. First, it is concluded that the set of rubrics was used as intended by the students and the teachers. Second, 12 of the 19 rubrics were feasible for the students to self-evaluate the ARV in inquiries that they conducted. Third, it appeared that novices should evaluate their inquiry plan and the completed inquiry as the two major parts of a self-evaluation of ARV in an inquiry. Moreover, the large number of rubrics hindered the students from developing a 'helicopter' view. Reflection on these findings led to the adoption of 21 items of the CoE model that seemed to be important for self-evaluation. When these 21 items were related to the two major parts for novices in a self-evaluation of ARV in an inquiry, it appeared that 13 CoE items needed to be worked out in rubrics. Moreover, to cover all items two extra tools should be designed: a holistic ARV card and an easy-to-use checklist. Fourth, the description of the benchmark samples required one topic instead of various topics. Last, there was no need to supplement the four design characteristics.

This chapter is based on the article: S.A.W. van der Jagt, L. van Rens, H.H. Schalk, A. Pilot, & J.J. Beishuizen. *A set of rubrics in learning to inquire by pre-university science students.* In preparation.

3.1 INTRODUCTION

At pre-university science level, learning to inquire has become an increasingly important part of the science education curriculum over the last decades (Abd-El-Khalick et al., 2004), including in the new Dutch formal science curriculum. Inquiries in school science subjects can have three main objectives. First, students develop knowledge about the natural world. Second, students learn how to use scientific equipment and how to improve their practical skills. Third, as a part of improving their procedural understanding, students learn how to evaluate the accuracy, reliability and validity (ARV) of the inquiries they conduct (Gott & Duggan, 1995; Millar, 2010). The meaning of the concepts *learning to inquire, accuracy, reliability and validity* in this thesis and previous research on these concepts are described extensively in section 1.3 (see Chapter 1).

This third objective can perhaps be achieved by showing pre-university science students the cognitive processes of scientists in authentic inquiries (Chinn & Malhotra, 2002; Van Rens et al., 2011). However, for these students it is already difficult to understand what is meant by evaluating the ARV of an inquiry, because they are novices at evaluating these aspects. In inquiry assignments, evaluating ARV does not often come into focus and as a result it is complicated for pre-university science students to improve their procedural understanding of ARV when they prepare, conduct and evaluate an inquiry (Lunetta, Hofstein, & Clough, 2007; Millar, 2010).

Flexible application of the evaluation of ARV in inquiries in different school science subjects is even more difficult for pre-university science students, despite the similarities in evaluating ARV in these different subjects (Roberts & Gott, 2002). Flexibility may be improved when students actively monitor their inquiries and judge their performances. This monitoring requires students to evaluate strategies and receive appropriate feedback more than once (Bransford, 2000).

Pre-university science students are novices in the domain of learning to inquire, and in evaluating the ARV of an inquiry. Bransford (2000) argued that such students should be provided with learning experiences in which they can recognise patterns in inquiries and in which they are supported in organising new information and its connection to their pre-existing knowledge. In organising new information for novices, self-evaluation can have a useful supportive function. Andrade and Valtcheva (2009) proved that self-evaluation involves both reflection on the task and revision of the work by students. Self-evaluation helped the students to focus on the main aspects of their task and to recognise the strengths and limitations of their work.

A possibility for pre-university science students in learning how to self-evaluate the ARV of an inquiry is to provide them with a coherent set of rubrics during an inquiry.

Sevian and Gonsalves (2008), for example, conducted a study on a rubric to examine to what extent it was helpful as an instrument for self-evaluating of scientific explanations. Rubrics support learning by making performance criteria explicit. These criteria make it easier for peers and the teacher to give feedback when students perform a self-evaluation of their work with rubrics (Jonsson & Svingby, 2007). Rubrics can be used as a formative instrument with qualitative descriptions of (levels of) performance criteria. However, many rubrics for secondary and higher education consist of ambiguous descriptions of performance levels for skills and strategies across their scale levels (Tierney & Simon, 2004). The review study of Jonsson and Svingby (2007) showed that most rubrics focus on the assessment of the content of student products (essays, reports) rather than on the processes or strategies of students. In particular, it is not known which characteristics are essential to design a set of rubrics that is feasible for pre-university science students in evaluating ARV during the enactment of inquiries. Therefore, the aim of this study is to gain more insight into the design characteristics for such a set of rubrics.

3.2 THEORETICAL PERSPECTIVE ON DESIGN CHARACTERISTICS OF RUBRICS

In order to design a set of rubrics with which pre-university science students can self-evaluate the ARV in successive science inquiry units, firstly four general aspects of rubrics for the evaluation of students' performances were identified in the literature and supplemented for the aim of this study. These four aspects are:

- *content* to self-evaluate ARV of inquiries,
- *extent of complexity* of the descriptions of the rubrics,
- *extent of detail* in view of the use by novices,
- *extent of general application* in different school science subjects.

These four general aspects will be described in more detail in this section.

3.2.1 Content

The first design characteristic concerns the content of the set of rubrics. Recent research in chemistry education (Van Rens, Pilot, & Van der Schee, 2010) and biology education (Schalk, Van der Schee, & Boersma, 2009) has shown that the use of the concepts of evidence (CoE) model (Gott, Duggan, Roberts, & Hussain, n.d.) can improve students' procedural understanding, including the ensuring of the ARV of an inquiry (Gott & Duggan, 2003). This suggests that the content of the rubrics with which students can self-evaluate the ARV during the enactment of an inquiry process can be related to the CoE.

Gott et al. (n.d) described 82 concepts of evidence in the CoE model. From this model 23 items were selected, because together these cover the CoE that were expected to be appropriate for evaluation of the ARV by pre-university science students during the enactment of successive science inquiry units. This selection of 23 items was based on the outcomes of the explorative study on the knowledge of pre-university students about ARV in inquiries (see Chapter 2). Moreover, these items were expected to be appropriate for a set of rubrics by which students can evaluate four main stages of conducting an inquiry: after they have formulated their inquiry plan, collected the data, analysed the data and completed the inquiry. Each of the 23 items was intended to let the students self-evaluate accuracy or reliability or validity in these stages while completing an inquiry process. Each item or combination of items can be elaborated in a rubric. All the 23 items together build up a set of 19 rubrics that students can use to evaluate ARV when they enact inquiries (see Table 3.1). Therefore, design characteristic 1, regarding the content of the instrument, was formulated as: the content of the set of rubrics is based on the CoE model.

3.2.2 Extent of complexity

The second design characteristic dealt with how to show the complexity of the 23 selected items in the different rubrics so that the students can self-evaluate the main stages of an inquiry process. According to Jonsson and Svingby (2007), rubrics can be described in performance levels so that the students can get a good orientation on the (expected) level of achievement. In many rubrics this description is done by first formulating the novice and expert levels, after which the criteria 'in between' are created, using wordings such as 'you are performing almost as described on the expert level'. These statements hardly give students any insight into how to improve their performance (Jonsson & Svingby, 2007; Mertler, 2001; Moskal, 2000). Furthermore, to have a stimulating effect on student learning in evaluating the ARV of an inquiry, the descriptions of the different levels of complexity in each rubric should be easy for the students to distinguish. To show the successive steps in evaluating the ARV of an inquiry, all descriptions in the rubrics should be represented hierarchically (Arter & McTighe, 2001; Moskal, 2000). This implies that a taxonomy is needed that describes the levels of complexity in a more sophisticated and hierarchical way. Chan, Tsui, Chan and Hong (2002) explored the application of three educational taxonomies – the Structure of Observed Learning Outcomes (SOLO) taxonomy, Bloom's taxonomy and a reflective thinking measurement model – in measuring students' cognitive learning outcomes. They concluded that the SOLO taxonomy is preferable for different kinds of learning outcomes. As such, the SOLO taxonomy was deemed suitable for this study, because it focuses on the levels of intended learning outcomes and is supportive for students in evaluating their performance at particular points in a learning task (Biggs & Tang, 2007; Hodges & Harvey, 2003; Levins & Pegg, 1993).

The SOLO taxonomy consists of five levels that increase in complexity: prestructural, unistructural, multistructural, relational and extended abstract (Chan et al., 2002). In our study, the prestructural level was defined as using the concepts of ARV in everyday language, or referring to daily life situations. Hence, students use tautology to

Table 3.1
Overview of the 19 rubrics for students to evaluate ARV in inquiries, the intended evaluation aspect and the corresponding 23 items from the CoE model

Rubric in stages of inquiry process	**Intended evaluation aspect**	**Item(s) from the CoE model**
After preparing the inquiry plan		
Theoretical framework	validity	Logical reasoning in theoretical framework
Inquiry question	validity	Specific and concrete inquiry question
Hypothesis	validity	Hypothesis can be tested by inquiry method
Inquiry method	reliability	Conduct control experiment
Drawing a sample	reliability	- Sample is sufficiently large - Sample is sufficiently varied
Preparation of tables to note down data	validity	Tables and graphs to sum up the results
Preparation of handling and analysis of data	accuracy	Measurement apparatus is sufficiently accurate
After collecting the data		
Experiment: independent variable	validity	The same independent variable throughout
Experiment: dependent variable	reliability	Reduce influence of other variables
Performing observations	accuracy	- Measure or observe with more than one observer - Measure or observe in an objective way - Measure or observe in a systematic way
Mean of and deviation in measurements	reliability	Repeat measurements, calculate averages and deviations
After handling the data		
Handling of outliers within measurements	accuracy	Measuring apparatus is calibrated
Comparability of results	reliability	Compare results with other inquiries
Drawing conclusion and use of evidence	validity	- Conclusion is based on inquiry results - Inquiry question is answered in the conclusion
Defining of patterns in results	accuracy	Measuring instrument has an adequate range
After completing the inquiry		
Evaluation of accuracy of the measurements	validity	Inquiry method fits inquiry question
Evaluation of reliability of the results	validity	Sufficient results to infer conclusion
Evaluation of validity of the conclusion	validity	Conclusion fully fits the inquiry question and inquiry method
Recommendation for supplementary inquiries	validity	Give recommendations for further inquiries

cover lack of understanding of ARV. The unistructural level was defined as using only one relevant aspect that is mostly based on quoting or memorisation, whereas at the multistructural level students use various relevant aspects but ignore any inconsistencies or relations between these aspects. At the relational level, the students use the inconsistencies or relations, but come to a firm conclusion, whereas the extended abstract level students recognise that new hypotheses can occur and that their inquiry is an example of a more general case. Based on the theory behind the SOLO taxonomy, the multistructural, relational and extended abstract levels of a rubric needed to be hierarchically built on the unistructural level.

When the SOLO taxonomy is properly applied to the content of each rubric in the set of rubrics, a student can only reach the relational level when the multistructural level is met completely (Biggs & Tang, 2007). Hence, design characteristic 2, on the design of feasible descriptions of intended complexity of the CoE items, was made operational by using the SOLO taxonomy to describe the levels of performance in each of the rubrics.

3.2.3 Extent of detail

The third design characteristic involved the extent of detail that is included in the set of rubrics. Rubrics can represent levels of performance in a holistic or an analytic way (Arter & McTighe, 2001; Mertler, 2001). A holistic rubric is used to make an overall judgement about the quality of a task, whereas analytic rubrics are used to evaluate different, smaller components in a task. Analytic rubrics are also useful in giving specific feedback to students and for self-evaluation purposes (Arter & McTighe, 2001). Especially students with less experience in performing a specific task, in our case self-evaluating the ARV of an inquiry, learn more from using rubrics with an analytical character than from using a holistic rubric (Chan et al., 2002). Therefore, for the purpose of this study, we opted for a set of analytic rubrics by which students learn in detail how to evaluate the ARV in different stages of an inquiry. The rubrics should thus give details to the students as novices and should support students with specific feedback. Hence, design characteristic 3, on the extent of detail, was formulated as: the set contains analytic rubrics that are feasible to be used by students in self-evaluating the four main stages of an inquiry.

3.2.4 Extent of general application

The fourth design characteristic concerned the function of the set of rubrics. Rubrics can have a specific function, for example, the evaluation of components of a single inquiry task (‘task specific’), or can have a general function for the evaluation of the same components in various inquiry tasks (‘generic’). Generic rubrics can be used across analogous tasks, e.g. all inquiry tasks in science subjects (Arter & McTighe, 2001; Jonsson & Svingby, 2007). The goal of this study was to let students self-evaluate inquiry tasks in different school science subjects, hence a set with generic rubrics was considered more suitable than task-specific ones. This implies, as Jonsson and Svingby (2007) argued, that besides a description of the SOLO taxonomy levels of complexity in each rubric, benchmark samples for each of the level descriptions in a rubric were needed. Benchmark samples in rubrics help students to interpret the generic descriptions in the rubrics as is intended. Jonsson and Svingby (2007) recommended that benchmark samples should be chosen with a variety as wide as the tasks wherein the rubrics are used. Hence, for the rubrics in this study the benchmark samples should be related to the content of the school science inquiry units in which the set of rubrics will be used. Therefore, design characteristic 4, on the generality of the set of rubrics, was formulated as: the rubrics consist of generic descriptions that are flexibly applicable in various inquiries and include supporting benchmark samples with a variety as wide as the inquiry tasks. In Table 3.2, an example is given of a rubric on the validity of the inquiry question, one of the 23 items (specific and concrete inquiry question) from the CoE model, with the five SOLO taxonomy levels of complexity and its benchmark samples.

Table 3.2
Rubric on the validity of an inquiry question: Example of one of the 19 rubrics

VALIDITY OF AN INQUIRY QUESTION

Ranking ↓ *Circle the description that best fits your inquiry question*	**Description**	**Benchmark sample**
1	You formulated the inquiry question on knowledge from your daily life.	*What is liquor?*
2	You formulated one variable in the inquiry question.	*What happens to your heart rate when you're standing upside down?*
3	You formulated the independent and dependent variables in the inquiry question.	*Which washing-up liquid cleans the best: one with zeolite or one with phosphates?*
4	You formulated the independent and the dependent variables in the inquiry question. The formulation also shows that you know how this inquiry fits into the research field.	*What is the relation between the angle of incidence of a laser in liquid and the angle of refraction of this laser?*
5	You formulated the independent and the dependent variables in the inquiry question. The formulation also shows that you understand how your inquiry relates to scientific claims about a similar issue.	*To what extent can rape oil be used to make a fuel that has the same calorific value as diesel oil?*

This study focused on whether the above described four design characteristics are sufficient and essential for the design and feasibility of a student self-evaluation instrument with rubrics. Therefore, the research question was:
To what extent are the design characteristics essential and sufficient for designing a set of rubrics that is feasible for pre-university science students to self-evaluate the ARV in successive science inquiry units?

3.3 METHOD

To design the set of rubrics the design research methodology as described by Gravemeijer and Cobb (2006) was used with a design phase, a test phase and retrospective analysis in two micro cycles of developing a local instruction theory.

In the first micro cycle, a prototype of the 19 rubrics, based on 23 items from the CoE model and the SOLO taxonomy, was studied for its compliance with the functions of a set of rubrics in evaluating the ARV of an inquiry. For this purpose, 16 science teachers and 22 pre-service science teachers used the draft set of rubrics to evaluate the ARV of students' inquiries. Their results were used to determine whether the rubrics covered the expected complexity in the students' inquiries. Reflection on the results was used for improving the descriptions in the rubrics and for incorporating the benchmark samples. Then the feasibility of the set of rubrics was extensively discussed in a group of three qualified and experienced pre-university teachers – of chemistry, biology and physics – and three educational science researchers. They discussed whether the set of rubrics contains the necessary items, covers the various school science units and covers the main stages in a school science inquiry unit. Adaptations to the set of rubrics were only made when consensus was reached. In addition, two pre-university science students commented on the language used in each rubric to make sure that they and their peers could understand the intended meaning of any description and of the benchmark samples. Their comments were mostly related to the level of difficulty of the terminology used and led to small adaptations. Compliance with all functions of the set of rubrics was fully met during the design process.

The second micro cycle, described in the next sections of this chapter, determined whether the set of rubrics that was designed in the first micro cycle is feasible for pre-university science students to evaluate ARV in inquiries that they conducted. This feasibility test was conducted using a qualitative research method (Cohen & Manion, 1994) with triangulation of data (Yin, 2003). This method was chosen because the feasibility of the set of rubrics needs to be tested in a naturally occurring setting of students in class (Collins, Joseph, & Bielaczyc, 2004). Student groups as well as individual students were taken as the unit of analysis (Cole & Engeström, 1993). To test whether the set of rubrics is feasible for student self-evaluation of ARV in inquiries that they perform in class, data were collected to find out whether: (a) each rubric is used as intended, (b) the students and teacher can work with all the rubrics, (c) the rubrics support the determination of the level of student understanding, and (d) the rubrics lead to a positive change in use of 'scientific' terminology by students and teachers (Nieveen, 2009). For an overview of the criteria for the feasibility of the set of rubrics for student self-evaluation and their connection with the four design characteristics see Table 3.3.

Table 3.3
Overview of criteria for the feasibility of the set of rubrics and their connection (√) with the design characteristics

Criteria for feasibility of the set of rubrics	**Design characteristic**			
	1	**2**	**3**	**4**
	content	**complexity**	**details**	**generality**
Support of intended self-evaluation of ARV				
a. used as intended		√	√	
b. the students and teachers can work with the rubrics	√	√	√	
c. determines level of student understanding		√	√	√
d. use of 'scientific' terminology	√		√	√

3.3.1 Participants

The participants in the second micro cycle were 24 pre-university science students (aged 16 or 17) and two pre-university science teachers from one upper secondary school in the Netherlands. These lessons were given as part of the 'free choice programme' of the students: they had to follow lessons in a school science subject of their own choice and this research project was one of these choices. The students and teachers did not receive an incentive for participating in this study, because the lessons took place during regular school periods.

The students were selected and invited by their science teachers to participate in the lessons of this research project. The first criterion for inclusion in the selection was that a student had to study biology, physics and chemistry at pre-university level. The second criterion for inclusion was that the students were motivated to perform inquiries in the school science subjects. During the selection procedure, no attention was paid to the previously achieved marks of the students in the school science subjects or whether they were talented in performing inquires. All selected students were used to doing practical work in their regular science classes, but had no experience yet in self-evaluating ARV while working on an inquiry. The teaching time for one inquiry unit was 160 minutes, which took place during regular school periods.

The two science teachers in biology and physics voluntarily agreed to participate in the study. One of them learned about this research during an oral presentation at a meeting of science teachers at university and informed his colleague about the possibility of participating in this research study. They both were (and are) motivated to improve the quality of students' inquiries in school science subjects. Both teachers have a master's degree and a first-grade teaching qualification. They had at least five years of teaching experience and were interested in students' inquiry projects as well as in working with a set of rubrics. To inform the teachers about the designed set of rubrics, both teachers came to a workshop of three hours that was arranged by the researcher. In this workshop, the student materials for the three inquiry units and the use of the rubrics were discussed and agreed upon.

3.3.2 Materials and procedure

In twelve pairs, the students conducted three inquiry units in general science, biology and physics in which the set of rubrics was implemented. The three inquiry units were taught in a one-year course. Each inquiry unit was composed of three successive lessons each lasting 160 minutes, so all the tasks of one inquiry unit together took a total of three times 160 minutes.

The designed teaching-learning process contained three successive inquiry units – general science, biology and physics. The students were expected to work with one or two other students on open inquiries during the inquiry units, so that they could discuss the various tasks involved in the units. For each unit, a workbook with tasks for the students and a teacher's guide was written. The teacher's guide contained a description of the intended functions of the set of rubrics and the intended teacher instructions during the completion of the inquiry units.

The first time the students had to use the set of rubrics, the teacher had a class discussion on why and how the students should use it. This discussion was followed by the students individually filling out the set of rubrics while evaluating the ARV of a previously performed students' inquiry about the influence of different drinks on the enamel of teeth. The filled-out rubrics were used to discuss which CoE was focused on in each rubric and how these CoE could contribute to the ARV of an inquiry. All the other times the students used a rubric, the teacher guided its use and talked with individual groups about how to apply the content of the rubrics in self-evaluating the ARV of their inquiries[3].

[3] *Writing the inquiry plan of the biology unit was done in class as well as the handling of the data and completing the inquiry. The accompanying rubrics were also filled out in class. At the zoo, after the students conducted the observations on animal behaviour, the four relevant rubrics about collecting data were filled out by all students, video-recorded by the researcher.*

Each inquiry unit was constructed around an inquiry task which the students planned, conducted and reflected on the process and outcomes. In the first inquiry unit, general science, the object of the inquiry task was a problem about the cooling rate of hot coffee. Then, in the biology unit, the inquiry task focused on the behaviour of animals in the zoo. Last, in the physics unit, the inquiry task regarded improving the traffic situation at a dangerous crossing by measuring the speed and braking distances of cyclists. In all the inquiry tasks, the students were given a problem that had to be approached by performing an inquiry. The students first had to design an inquiry plan. By means of instructions in the workbooks, the students were asked to formulate an inquiry question and a hypothesis and to design an inquiry method. They then did a pilot experiment to test the practicability of their described inquiry method. Using the results and experiences of the pilot experiment, the students were asked to adjust, if necessary, their inquiry plans. Next, the students were expected to use the relevant rubrics to self-evaluate the ARV of their inquiry plan and to decide whether the plan needed any changes. Then, the students had to conduct the experiments according to their plans to collect data and self-evaluate the ARV of the data collection with the relevant rubrics. After this, they had to work out the obtained data into conveniently arranged results, and draw conclusions. Afterwards, the students were expected to self-evaluate the ARV of their completed inquiries with the relevant rubrics and were asked to write down their suggestions for enhancing the ARV in future inquiries.

In all the inquiry units, students' self-evaluation of the ARV of their inquiries using the rubrics was intended to take place (i) after preparing an inquiry plan, (ii) after collecting their data, (iii) after handling the data, and (iv) after completing the inquiry. The student workbooks contained instructions for the students to ask their teacher for the relevant rubrics at the intended points during the teaching-learning process.

To guide and support the students' performance in each inquiry task in the teaching-learning process, the inquiry units contained three types of supporting tasks:

i) Orientation tasks, on the evaluation of ARV with the rubrics and on the meaning of the 23 CoE items in the set of rubrics;
ii) Peer feedback with the set of rubrics;
iii) Flexible application of evaluation of ARV with the set of rubrics from one inquiry to another.

Directly after each lesson, two researchers independently scored the students' written inquiry plans, collected data, handled data and/or completed inquiries. This scoring was done with the same rubrics as the ones that were used by the student groups. The proportion of agreement of the scores of both researchers was 73%. The inconsistencies between scores were discussed until consensus was reached (Janesick, 2000). The completed rubrics from the researchers, with consensus about the score, were given to the student groups and the teacher at the start of the next period. This gave the students an opportunity to compare their own rubric scores with those of the researchers. The students also had the opportunity to ask the teacher and researchers questions about dissimilarities and uncertainties so that they could improve their understanding about the rubrics' content and function. After the planning phase, the teacher also asked the students – if necessary– to adapt their inquiry plan.

3.3.3 Data collection

All the lessons in which the student groups used rubrics were observed and video-recorded by one of the researchers. The observations and crucial moments in the videos of each lesson were worked out in a field report that was read by the teachers and discussed until consensus was reached with the teachers about the meaning of their statements. Four groups of students were audio-recorded during all periods. All rubrics that the twelve student groups and the researchers completed were collected together with the students' worksheets with inquiry plans, their research data and results, their conclusions and their general opinions on the ARV in the enacted inquiry. All the students' and researchers' notes on rubrics were also collected.

Immediately after finishing an inquiry unit, all the students individually completed a questionnaire. The questions in the questionnaire were:

- *Was it possible to evaluate the quality of your inquiry with the rubrics? Mention strong and weak points.*
- *What could be improved?*
- *Did you use all rubrics? If no, give a reason.*
- *Do you have any other remarks regarding the use of the rubrics?*

After the students had completed the questionnaires, four students volunteered for an in-depth interview related to the questions in the questionnaire. Each interview took about 30 minutes. All interviews were audio-recorded and transcribed.

The teachers were interviewed immediately after the lessons in which the students self-evaluated one of the main stages in each inquiry unit. Each interview took about 20 minutes. The interview questions were:

- *Was it possible for the students to self-evaluate the quality of their inquiry with the rubrics? What are the strong and weak points in the rubrics, thinking of the descriptions, benchmark samples and use in class?*
- *What could be improved?*

All interviews were audio-recorded and transcribed. For the purpose of this thesis, quotes of students and teachers were translated from Dutch into English.

3.3.4 Data analysis

The data sources were analysed using four criteria as to whether the set of rubrics has the potential to support the student self-evaluation of ARV in inquiries (see Table 3.3). These criteria were chosen because of their expected contribution to the feasibility of the designed set of rubrics for self-evaluating the ARV in inquiries in different school science subjects.

Criterion a was whether the students used the rubrics as intended. This was expected to be visible in which rubrics students used and did not use in the different inquiry stages in the successive science inquiry units. The number of groups that used a rubric for self-evaluation of their inquiry plan, their data collection, their data handling and their completed inquiry in each of the three successive inquiry units was determined. A rubric was considered as feasible for self-evaluation when at least six groups (50%) of students used it. This analysis was supplemented by observations on the support and encouragement given by the teachers on the use of the rubrics by the student groups and observations on the use of and questions about the rubrics from students.

Criterion b was whether the students and teachers could work with the set of rubrics. The students' responses in the questionnaires, the transcripts and their notes on the rubrics were analysed on: (i) the number of rubrics, and (ii) the clarity of the descriptions and benchmark samples. Responses from the teachers' interviews, data from the field reports and the researchers' notes on the rubrics were analysed on: (i) the size of the set of rubrics, and (ii) the hierarchy of the descriptions and the feasibility of the benchmark samples in the various rubrics.

Criterion c involved the support in determining the level of student understanding of evaluating the ARV of an inquiry. In the analysis, the rankings in the rubrics that were used in the successive inquiry units by an average of ten or more student groups (83%) were compared with the rankings of the researchers. The degree of correspondence between the rankings from the students and the researchers was used to determine whether the rubrics were feasible to determine the level of student understanding of ARV and the support students needed to improve this understanding by using the rubrics. As Hafner and Hafner (2003) argued: the degree of correspondence between students' rankings and the researchers' rankings is an indicator of the support that students need from the teacher in a specific task.

Criterion d concerned whether the use of the rubrics did increase the use of CoE in the conversations between students and teachers. The students' worksheets, the transcripts of students' and teachers' conversations from the audio recordings in class and the interviews were analysed on the occurrence of appropriate CoE about self-evaluation of ARV that could be related to the set of rubrics.

3.4 FINDINGS

3.4.1 Actual use of the set of rubrics

The findings regarding criterion a, what rubrics the students actually use to self-evaluate the four main stages in the three successive science inquiry units, are presented in Table 3.4. On average, 12 rubrics were used by six or more of the student groups in the three inquiry units. These 12 rubrics were: Inquiry question, Comparability of results, Hypothesis, Drawing conclusion and use of evidence, Evaluation of accuracy of the measurements, Evaluation of reliability of the results, Evaluation of validity of the conclusion, Theoretical framework, Inquiry method, Experiment: independent variable, Experiment: dependent variable and Recommendation for supplementary inquiries. On average, seven rubrics were completed by five groups or fewer: Mean and deviation in measurements, Drawing a sample, Handling of outliers within measurements, Preparation of tables to note down data, Preparation of handling and analysis of data, Performing observations and Defining of patterns in results.

Analysis of the field reports regarding the teachers' instructions about the use of the set of rubrics showed that the teacher arranged a whole-class discussion when introducing the rubrics to self-evaluate the ARV of an inquiry during the general science unit. All of the students were given a set of rubrics before this instruction and completed these directly afterwards to self-evaluate their inquiry plans. During the general science inquiry unit, the teacher encouraged all student groups to complete the set of rubrics. During the first of the three biology periods, the teacher again gave a whole-class instruction to the students: 'Don't forget to complete the rubrics for the research project of [name of researcher] from the university'. At other times during the biology unit, half of the student groups asked the teacher for their sets of rubrics after they had read the instruction in their worksheets to ask for their rubrics. Students who did not ask for rubrics received a copy at the end of the period when the teacher inspected whether the students had fully completed their inquiry. The six student groups that completed the rubrics at the end of the three periods appeared to put rankings without carefully reading the description and benchmark samples in the rubrics. During the physics unit, the teacher handed out the relevant rubrics at the beginning of each period and gave a whole-class instruction to complete the rubrics at 'the points indicated during the inquiry'. About half of the students completed the rubrics when this action was pointed out in their worksheets. The teacher did not further encourage the individual student groups to complete the rubrics.

Analysis of the worksheets and the field reports regarding the use of rubrics showed that none of the student groups improved their inquiry plan after ranking their plans with rubrics and after receiving the researchers' rankings in the rubrics, although this was explicitly requested by the teacher. During the third period of the general science inquiry unit, one student group asked the researcher about the comparison between their own ranking and the ranking of the researchers. During the other periods, the students did not ask questions about this comparison. Analysis of the field reports shows that the students did not ask questions about uncertainties in the descriptions in the rubrics.

Table 3.4
Number of student groups (n=12 per unit) that actually used a rubric in the successive inquiry units, by percentage of use

Rubric	General science	Biology	Physics	Percentage student groups
Inquiry question	12	11	12	97%
Comparability of results	11	12	11	94%
Hypothesis	12	12	10	94%
Drawing conclusion and use of evidence	12	11	10	92%
Evaluation of accuracy of the measurements	12	10	10	89%
Evaluation of reliability of the results	12	10	10	89%
Evaluation of validity of the conclusion	12	9	9	83%
Theoretical framework	*	10	9	79%
Inquiry method	12	9	6	75%
Experiment: independent variable	10	*	8	75%
Experiment: dependent variable	9	*	7	67%
Recommendation for supplementary inquiries	7	5	6	50%
Means & deviations in measurements	5	*	4	38%
Drawing a sample	*	4	*	33%
Handling of outliers within measurements	4	*	2	25%
Preparation of tables to note down data	*	2	2	17%
Preparation of handling and analysis of data	3	*	1	17%
Performing observations	*	2	*	17%
Defining of patterns in results	0	*	0	0%

Note. *: not intended to be used by the students in this inquiry unit.*

3.4.2 Experience of use of the set of rubrics

Analysis of the students' responses in the questionnaires, interviews and their written notes on the completed rubrics included, in total, 73 statements about experiences with the use of the rubrics (criterion b). Thirty-five student statements (49%) were about the size of the set of rubrics, descriptions and benchmark samples, which are described below in more detail. Furthermore, 21 statements (29%) were about the use of the rubrics in general, such as: *'It was good, because I had the chance to improve* *my inquiry.'* and *'I did not like working with rubrics.'* Eight student statements (11%) were about the instruction by the teachers. Nine statements (12%) were too ambiguous to categorise, like *'Interesting'*.

The transcripts of the teachers' interviews, the field reports and the researchers' notes included in total 41 statements about their experiences with the set of rubrics. Of these, 25 statements were about the size of the set and the descriptions and benchmark samples, which are described in more detail below. Furthermore, 10 general statements (24%) were about the use of the rubrics by students such as:

- *The rubrics motivate students to make improvements in parts of their inquiries.*
- *The students had not thought about the ARV of an inquiry before starting with the general science inquiry unit.*

The teachers made 6 statements (15%) on the issue of spending more effort on classroom management than on helping individual students with the rubrics as they had planned.

(i) Size of the set of rubrics

Further analysis of the responses in the students' questionnaires and interviews showed six responses (8%) with positive connotations and eleven responses (15%) with negative connotations regarding the size of the set of rubrics. Examples of student quotes are:

- *It was an easy-reference. I could see exactly what was good and what could be improved.*
- *It took too long to complete everything and it became dull to repeat it every time.*
- *Half of the rubrics I did not use, because I had not done these things during my inquiry.*

Further analysis of the teachers' responses in the interviews showed that, regarding the size of the set, the teachers together noticed three times (13%) that it took a lot of time for the students to complete the rubrics. However, they also stated that they expected that a more frequent use of rubrics by the students would remedy this problem. Analysis of the field reports and the researchers' notes on the rubrics showed that they experience that the time needed to complete the rubrics decreases during the successive inquiry units. Furthermore, the researchers made 7 remarks about the need for additions when the students did not describe the required information about their inquiry on their worksheets. As an example, one comment involving the rubric 'Performing observations' is:

- *They only wrote down: 'We will perform observations during our inquiry', but did not write how they would perform the observations, what the objects of their observations are, etcetera. Actually, is it necessary to guide novices in this rubric better?*

Analysis of the field reports and the completed worksheets showed that students had difficulties giving a general judgement about the ARV of their completed inquiries and scarcely used the relevant rubrics for doing this.

(ii) Descriptions, benchmark samples and hierarchy

Analysis of the student responses in the questionnaires and the interviews revealed two responses (3%) with positive connotations and seven (10%) with negative connotations regarding the clarity of the language of the rubrics and the benchmark samples. The quotes with positive connotations about the descriptions and benchmark samples are:

- *Most of the time I understood what the difference was between the descriptions, and sometimes I completed what matched my inquiry best, and neglected some [parts] of the description. It was mostly elucidated by the benchmark samples.*
- *The benchmark samples were useful in understanding the descriptions [in the rubric] and did help to check your own work.*

Quotes with negative connotations include:

- *The rubrics are handy, but the descriptions in the various levels were quite difficult to understand. That needs improvement.*
- *[It was] not always clear on which [level] I had performed. It fitted better in between levels than in a specific one or I made a mix of parts of different levels.*
- *Sometimes I did not see why one benchmark sample was better than another was. I thought they might be reversed by accident.*

The teachers' responses in the interviews contained four remarks (17%) about the descriptions and benchmark samples in the rubrics. One quote with positive and one with negative connotations are:

- *Students can make these improvements because the levels in the rubrics are hierarchically described and create room for student thinking and reflection.*
- *The wide variety in subjects makes it difficult to see the coherence between the benchmark samples in each rubric.*

Analysis of the researchers' notes about the feasibility of the rubrics showed that the researchers experienced problems with the hierarchy in 6 out of 19 rubrics: Research method, Preparation of tables, Experiment: independent variable, Experiment: dependent variable, Comparability of results and Evaluation of accuracy. As an example, this is a comment about the rubric 'Research method':

- *Based on some components of the inquiry plan, I should rank it at level 4, but it does not fully match the descriptions at the previous levels.*

3.4.3 Determination of support needed by students

The level of student understanding of ARV, and as a consequence the identification of the student groups that needed support from the teacher to improve their understanding (criterion c) is analysed for the seven rubrics that were completed by on average at least 10 (83%) student groups in the successive inquiry units (see Table 3.4). Table 3.5 presents the degree of correspondence between the student groups' and researchers' rankings in the rubrics in the three successive inquiry units. It shows that the correspondence is the highest in the rubrics 'Inquiry question' and 'Drawing conclusion and use of evidence'. The largest differences in the student groups' and researchers' rankings occur in the rubrics 'Hypothesis' and 'Evaluation of accuracy of the measurements'.

Table 3.5
Degree of correspondence (in %) between student groups' and researchers' rankings in rubrics as completed by on average at least 10 student groups in the successive inquiry units

	Degree of correspondence (%)		
Rubrics	**General science**	**Biology**	**Physics**
Inquiry question	83%	82%	83%
Hypothesis	25%	8%	20%
Comparability of results	55%	42%	36%
Drawing conclusion and use of evidence	83%	82%	90%
Evaluation of accuracy of the measurements	33%	17%	25%
Evaluation of reliability of the results	58%	67%	75%
Evaluation of validity of the conclusion	67%	56%	56%

3.4.4 The use of CoE by students

Analysis of the transcripts of the teachers' interviews and the field reports showed that, regarding criterion d, the biology teacher experienced during the biology inquiry unit that the students' discussions were more directed to the use of 'scientific' terminology:

- *Students talk more about their inquiry plan than in the previous inquiry unit, especially about ARV. Maybe, the reason is that in the general science unit they explicitly worked on those.*

Analysis of the field reports revealed that the teachers did not guide the students when they neglected to self-evaluate CoE like deviation and range. The teachers did not use rubrics to help students in self-evaluating the CoE that they neglected during the evaluation of ARV of inquiries. Moreover, the analysis showed that one of the teachers stated that students first have to know the meaning of a CoE before they can self-evaluate their performance regarding that CoE. As she said,

- *As long as students don't know what deviation is and how to determine the deviation in their measurements, then they won't make this step during an inquiry and can't evaluate their performance.*

This concerns an important aspect of the teaching-learning process, but does not influence the feasibility of the rubrics as such.

3.5 DISCUSSION AND CONCLUSION

In the first micro cycle, compliance with the functions of the set of rubrics was met fully during the design process. Testing the feasibility of the set of rubrics for the intended teaching-learning process (second micro cycle) was guided by four criteria. Reflection on the first criterion (a) shows that the rubrics are used as intended by the students and teachers. Twelve of the 19 rubrics are feasible for pre-university students as novices in self-evaluating the ARV of inquiries. These were used by most of the student groups when they self-evaluated the ARV in the main stages of an inquiry during their enactment of successive school science inquiry units (see Table 3.4). Seven rubrics are less feasible because only a few groups used them, among which the rubric 'Defining of patterns in results' was used by none of the groups. A reason for this might be that the items from the CoE model in these unused rubrics were, according to the students, not applicable in the inquiry process that they conducted, although the designers expected their application. Novices very often do not yet have the flexibility to see whether a concept (in this study the CoE 'Measuring instrument has an adequate range') fits in a part of the inquiry process (Bransford, 2000), which could mean that pre-university students should fully understand first what is described in a rubric before they can apply it to their own inquiry. Practising with one example of an inquiry, as was done in this study, did not seem enough for them to gain a complete understanding of the descriptions in the rubrics. Furthermore, rubrics that in this study were only completed by students in one inquiry unit, for example, the rubric 'Performing observations', could be integrated in the workbook. In this way, students can learn why the three corresponding CoE items – measure or observe with more than one observer, measure or observe in an objective way and measure or observe in a systematic way – are important for accuracy in a particular inquiry but not in all inquiries.

Reflection on the second criterion (b), whether the students and teachers can work with the set of rubrics, leads to a couple of possibilities for improving their feasibility. Regarding the size of the set of rubrics, the analysis of teachers' responses in the rubrics, as well as of the field reports, supports the conclusion that the number of rubrics should be reduced to avoid students getting lost. The analysis of the students' worksheets indicates that writing the inquiry plan and completing the inquiry are the major analytical parts for novices. The rubrics require a connection with these major parts. The size of the set of rubrics also seems to hinder students in developing a helicopter view in self-evaluating the ARV of the inquiry as a whole. Through the number of rubrics, these novices found it hard to detect the parts and issues they should identify to evaluate appropriately the inquiry as a whole. An option is to provide students with an overview tool and an easy-to-use checklist to control their performance before they apply a set of rubrics for evaluation purposes. Such a checklist seems to be helpful when novices have difficulties in using a self-evaluation instrument (Sadler, 1989). The checklist can help students to decide whether they first have to add or repeat parts of their inquiry before they can reflect appropriately on its ARV.

Regarding the descriptions and benchmark samples, indications have been found that some descriptions were not clearly formulated for students and therefore less feasible in self-evaluating the ARV of inquiries. These unclear formulated descriptions need minor revisions. A major revision should be made to the variety of the benchmark samples. In contrast to the recommendations of Jonsson and Svingby (2007), to use a variety of benchmark samples as broad as the range of school science subjects in which the rubrics will be used, we found that students were confused by the benchmark samples from different science domains and did not see the coherence between them. However, some students also mentioned that the benchmark samples elucidated the more general descriptions. This notion makes it worthwhile to maintain the benchmark samples in the rubrics, but relate all of them to the same scientific topic.

From the findings regarding the hierarchy of the descriptions in the rubrics, it can be concluded that the majority of the rubrics actual have multistructural, relational and extended abstract levels that are built in a hierarchical way on the unistructural level, in accordance with the SOLO taxonomy. However, analysis of the researchers' notes of the class observations shows that six rubrics appeared not to be hierarchically ordered. A more detailed analysis showed that the descriptions in the rubrics 'Experiment: independent variable', 'Experiment: dependent variable' and 'Comparability of results' were equal in complexity and did not comply with the hierarchical SOLO taxonomy. This lack of hierarchy in the descriptions can be expected to influence the feasibility of these rubrics for pre-university science students, because they do not understand 'where they head to'.

With regard to the third criterion (c), support in determining the level of student understanding, it can be concluded that the set of rubrics appeared to be feasible for determining which students need (more) individual support, for example, in formulating a valid hypothesis. Further research of the relevant rubrics is needed to analyse and evaluate this criterion in more detail. Because of the limited actual use in this study, only 7 of the 19 rubrics were used in determining the level of student understanding (see Table 3.5).

Reflecting on the last criterion (d), change in use of 'scientific' terminology by students and teachers, the findings from the teachers' interviews and field reports show that a few students started to use terminology that can be related to the use of the rubrics, perhaps as a consequence of using the set of rubrics. As one of the teachers mentioned, the students had not thought about the ARV of an inquiry before starting the general science inquiry unit. During the successive inquiry units, some of them started to refer to these aspects tentatively, which also may be due to the iterative use of the set of rubrics. It is expected that the use of scientific terminology can be further improved when the dialogue between students and teachers is more focused on the ARV of an inquiry and less on practical issues in conducting inquiries, for example, guided by a feasible set of rubrics. The two participating teachers in this study were not used to evaluating the ARV of inquiries with students, which could also have influenced the feasibility of the use of rubrics to evaluate ARV of inquiries in class. When teachers have difficulties in using the set of rubrics, it cannot be expected that students as novices can be taught how to use these rubrics appropriately. The teachers should be trained to become more acquainted with teaching and supervising these aspects of an inquiry (Oliviera, 2010; Van der Schee & Rijborz, 2003).

This study focused on the question: *To what extent are the design characteristics essential and sufficient for designing a set of rubrics that is feasible for pre-university science students to self-evaluate the ARV in successive science inquiry units?* Based on the outcomes of this study, it was concluded that the four design characteristics for a feasible set of rubrics for self-evaluating the ARV of inquiries had to be sharpened. These four extended and more detailed design characteristics were used in further studies on the design and implementation of a revised set of rubrics (see Chapters 4 to 6). The findings did not suggest adding or removing design characteristics. The conclusions and implications for the rubrics can be summarised as:

(i) Content

Based on the enactment in class, the focus of the evaluation of ARV by novices should be on evaluating the inquiry plan and the completed inquiry. In line with this, two CoE items – 'Tables and graphs to sum up the results' and 'Conclusion fully fits the inquiry question and inquiry method' – could be excluded. Hence, 21 of the 23 selected CoE items seemed to be important in self-evaluating the ARV in an inquiry by pre-university science students. To make the self-evaluation instrument more feasible, fewer rubrics, 13 items of the CoE model, were considered as major concepts in evaluating ARV in an inquiry. These concepts were related to the following rubrics:

- Theoretical framework
- Inquiry question
- Hypothesis
- Inquiry method
- Drawing a sample
- Means and deviations in measurements
- Drawing conclusions and use of evidence
- Evaluation of accuracy of the measurements
- Evaluation of reliability of the results
- Evaluation of validity of the conclusion
- Recommendations for supplementary inquiries

Although they were used by fewer than half of the student groups, the rubrics 'Drawing a sample', 'Means and deviations in measurements' and 'Recommendations for supplementary inquiries' should be included in a revised set of rubrics, because the four CoE items in these rubrics (see Table 3.1) were considered by the researchers as important concepts in evaluating the ARV in inquiries. Hence, 21 CoE items should be worked out in a self-evaluation instrument. Thirteen of the 21 items should occur in rubrics and the other 8 items from the CoE model, which were also valuable for novices in learning how to self-evaluate the ARV of inquiries, could be worked out in less complexity as in the set of rubrics was done. An option to improve the feasibility could be the design of an easy-to-use checklist for evaluating ARV in an inquiry.

(ii) Extent of complexity

Each rubric had five hierarchical levels in compliance with the SOLO taxonomy. Prestructural and unistructural levels were based on the prerequisite knowledge of pre-university students. Multistructural, relational and extended abstract levels were built in a hierarchical way on the unistructural level. Minor revisions to the formulation and the hierarchy of the descriptions in some rubrics were made in a revised set of rubrics (see Chapter 4).

(iii) Extent of detail

The rubrics in the set had an analytic trait. To improve the feasibility of the set of rubrics, it should be accompanied by a tool that gives a holistic overview or ARV card of the relation between the CoE as described in the different rubrics and the ARV of a completed inquiry.

(iv) Extent of general application

The set of rubrics was designed to be generic for self-evaluating the ARV in inquiry units in the different school science subjects. It included benchmark samples to elucidate the generic descriptions. To improve the feasibility of the set of rubrics, it was concluded that the benchmark samples should all refer to the same inquiry topic and be more univocal to the students.

In summary, a set of rubrics that is feasible for evaluating the ARV in inquiries in different school science subjects by pre-university science students could be designed with the use of the four design characteristics as described in this study. The set of rubrics proved to be largely feasible for teaching pre-university science students how to self-evaluate the ARV of an inquiry, but to improve the design it is necessary to reduce the number of rubrics and to work out 13 items from the CoE model in rubrics. In addition, to give the students a helicopter view, a holistic ARV card and an easy-to-use-checklist that cover all 21 items of the CoE model need to be designed. The feasibility of this revised and more extended self-evaluation instrument was investigated in the next study, as described in Chapter 4.

This study was based on data from three successive inquiry units in a classroom setting at one school. The feasibility of the set of rubrics in other inquiries and at other schools needed further study with pre-university science students. The retrospective analysis (Gravemeijer & Cobb, 2006) on the local instruction theory and the functions of the set of rubrics in the successive inquiry units also provided information for the revision of the rubrics and its implementation in class. A limitation of this study was that the participating teachers had little experience in supporting students in self-evaluating ARV, although they did attend a workshop on the subject. During the first research cycle misinterpretation of the use of the rubrics in the intended teaching-learning process often occurred, which could have influenced the outcomes of the first research cycle. However, the exploration of the set of rubrics in the successive inquiry units with the 24 students showed that it was feasible to a substantial extent for self-evaluation of ARV in inquiries by pre-university science students.

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.

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: Using performance criteria for assessing and improving student performance. Thousand Oaks, CA: Corwin Press.

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.

Chan, C.C., Tsui, M. S., Chan, M.Y.C., & Hong, J.H. (2002). Applying the Structure of the Observed Learning Outcomes (SOLO) taxonomy on students' learning outcomes: An empirical study. Assessment & Evaluation in Higher Education, 27(6), 511–527.

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.

Cohen, L., & Manion, L. (1994). Research methods in education (4th ed.). London: Routledge.

Cole, M., & Engeström, Y. (1993). A cultural-historical approach to distributed cognition. In G. Salomon (Ed.) Distributed cognitions. Psychological and educational considerations (pp. 1–46). Cambridge: Cambridge University Press.

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

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

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

Gott, R., & Duggan, S. (2003). Understanding and using scientific evidence: How to critically evaluate data. London: Sage Publications.

Gott, R., Duggan, S., Roberts, R., & Hussain, A. (n.d.). Research into understanding scientific evidence. Retrieved on 19 May 2012 from http://www.dur.ac.uk/rosalyn.roberts/Evidence/cofev.htm

Gravemeijer, K., & Cobb, P. (2006). Design research from a learning perspective. In J. Van den Akker, K. Gravemeijer, S. McKenney, & N. Nieveen (Eds.), Educational design research (pp. 17–51). London and New York: Routledge.

Hafner, J., & Hafner, P. (2003). Quantitative analysis of the rubric as an assessment tool: an empirical study of student peer group rating. International Journal of Science Education, 25(12), 1509–1528.

Hodges, L. C., & Harvey, L. C. (2003). Evaluation of student learning in organic chemistry using the SOLO taxonomy. Journal of Chemical Education, 80(7), 785–787.

Hodson, D. (1999). Building a case for a sociocultural and inquiry-oriented view of science education. Journal of Science Education and Technology, 8(3), 241–249.

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.

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 [electronic version]. Practical Assessment, Research & Evaluation, 7(25). Retrieved on 30 March 2009 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.

Moskal, B. M. (2000). Scoring rubrics: What, when and how? [electronic version]. Practical Assessment, Research & Evaluation, 7(3). Retrieved on 2 April 2009 from http://PAREonline.net/getvn.asp?v=7&n=3

Nieveen, N. (2009). Formative evaluation in educational design research. In T. Plomp & N. Nieveen (Eds), An introduction to educational design research. Enschede, The Netherlands: SLO.

Oliviera, A. W. (2010). Improving teacher questioning in science inquiry discussions through professional development. Journal of Research in Science Teaching, 47(4), 422–453.

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.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. Instructional Science, 18(2), 119–144.

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 & 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: Bèta Press.

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.

Tierney, R., & Simon, M. (2004). What's still wrong with rubrics: Focusing on the consistency of performance criteria across scale levels [electronic version]. Practical Assessment, Research & Evaluation, 9(2). Retrieved on 30 March 2009 from http://PAREonline.net/getvn.asp?v=9&n=2

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

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., Van Muijlwijk, J., Beishuizen, J., & Van der Schee, J. (2011). 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.