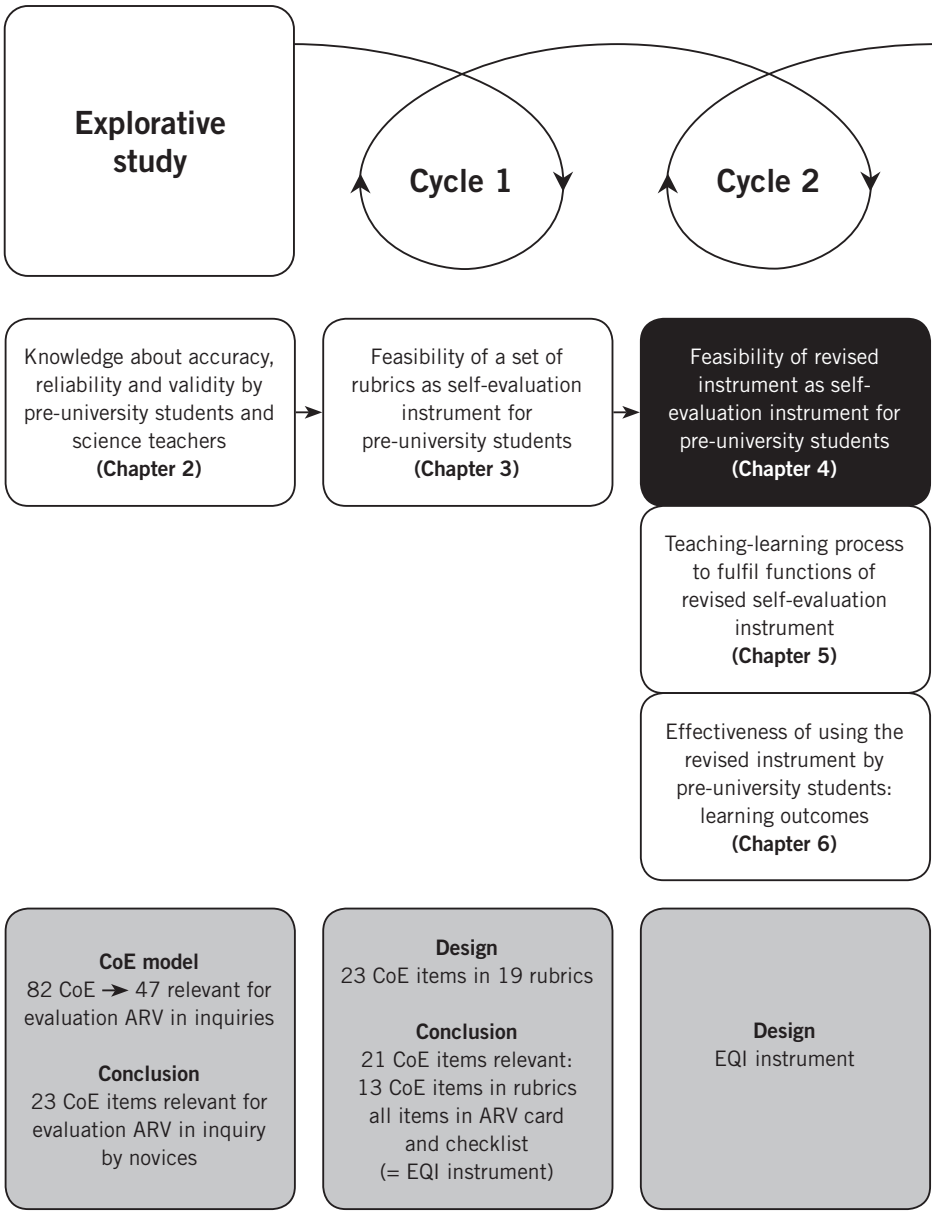

saskia_proefschrift_e_chapter_04.indd 76 27-10-16 10:49

FEASIBILITY OF REVISED INSTRUMENT AS SELF-EVALUATION INSTRUMENT FOR PRE-UNIVERSITY STUDENTS

CHAPTER 4

This study is based on design research and is focused on the feasibility of a revised self-evaluation instrument, the Evaluation of Quality of Inquiries (EQI) instrument, with which pre-university science students can evaluate the accuracy, reliability and validity (ARV) of inquiries. The instrument has four design characteristics regarding: the content (concepts of evidence (CoE) model), the extent of complexity (Structure of Observed Learning Outcomes (SOLO) taxonomy), the extent of detail and the general application, respectively. To determine its feasibility in class, 27 students used the instrument in three successive inquiry units. The instruments completed by students and teacher, lesson observations, audio and video recordings, student responses in interviews and questionnaires as well as reflection reports of the teacher were analysed. From this analysis, it was concluded that the criteria on 'the instrument is used as intended', and on 'students and teacher can work with the EQI', were fulfilled completely. The criteria 'the EQI instrument supports the determination of the level of student understanding' and on 'the EQI instrument leads to a positive change in scientific terminology by students and teacher' were sufficiently fulfilled. Hence, it is plausible that the EQI is feasible for pre-university science students in evaluating the ARV of inquiries.

This design research is part of one of the studies in the second research cycle and is reported in this chapter. The other two studies are respectively presented in Chapter 5 and 6.

This chapter is based on the article published as: Van der Jagt, S.A.W., Van Rens, L., Schalk, H.H., Pilot, A., & Beishuizen, J.J. (2013). Een instrument voor bovenbouw vwo-leerlingen om de kwaliteit van hun natuurwetenschappelijk onderzoek te evalueren [with summary in English]. *Pedagogische Studiën, 90(2),* pp. 47-62.

saskia_proefschrift_e_chapter_04.indd 77 27-10-16 10:49

4.1 INTRODUCTION

The Dutch pre-university science curriculum, in line with a worldwide trend, contains scientific inquiry learning for students. This curriculum involves the development of student insight into the nature of science as well as into the way scientists work and think (Aarsen & Van der Valk, 2008; Abd-El-Khalick et al., 2004; Van Rens, Van Muijlwijk, Beishuizen, & Van der Schee, 2011; Van Rens et al., 2014). Scientific inquiry learning in the school science subjects has various aims for the students. The emphasis can lie on an increase in scientific knowledge, practical skills or procedural understanding (Gott & Duggan, 1995; Millar, 2010), for example, evaluating accuracy, reliability and validity (ARV) in an inquiry.

Increasing students' procedural understanding helps them to apprehend the criteria that scientists use and their way of working when they plan, conduct and evaluate research (Chinn & Malhotra, 2002). However, pre-university students are novices in evaluating ARV in scientific inquiries. As shown in Chapter 2, it is often unclear to students what the meaning of ARV is in inquiries. Moreover, students at the pre-university level are mostly used to practical work, in which they use a stepwise approach without much ownership and often with little reflection on their actions (Lunetta, Hofstein, & Clough, 2007; Schalk, Van der Schee, & Boersma, 2013). Therefore, pre-university students hardly know how to evaluate ARV in practical tasks.

Furthermore, these students do not generally realise that in the different science subjects there is much conformity on how to enact and evaluate the ARV in inquiries (Roberts & Gott, 2002). Bransford (2000) suggested that the ability of students to transfer concepts from one inquiry to another can increase when students actively control their activities as well as evaluate the results in a structured way. Bransford (2000) concluded that novices in a certain domain, like pre-university students when evaluating the ARV of an inquiry, need to do learning activities in which they can recognise domain-specific patterns and in which they are supported in structuring 'new' knowledge and in linking this knowledge to their prior knowledge. Andrade and Valtcheva (2009) showed that the use of a self-evaluation instrument 'forces' students to reflect on and revise their actions. Self-evaluation focuses the students' attention on to the main, most important aspects of a task and also helps them in structuring 'new' knowledge as well as recognising the weaker and stronger parts of their performance.

A self-evaluation instrument probably also helps pre-university science students in learning how to evaluate the ARV in inquiries. Sevian and Gonsalves (2008) showed in their research on undergraduate students that a coherent set of rubrics can be used in science education. A rubric contains descriptions of performances of novices and

saskia_proefschrift_e_chapter_04.indd 78 27-10-16 10:49

experts with preferably one or more descriptions of the in-between levels. Each level description needs to be clearly distinguished from the other descriptions in a hierarchical way (Burke, 2006). If the possible actions of the students on the different levels are explicitly described, then the rubrics can be supportive in the student learning process. With explicit descriptions in rubrics, students can independently evaluate their actions and gain insight into how to improve (Jonsson & Svingby, 2007).

The review study of Tierney and Simon (2004) showed that sometimes rubrics designed for students to evaluate their inquiries do not offer sufficient support for novices in learning to evaluate the ARV in inquiries. Dutch pre-university science teachers experience the same problem (personal communication). These rubrics suffer from vagueness in descriptions, like: 'You found sufficient literature references that fit your inquiry question and sub questions'. From this description it is unclear for the user of the rubric how many references are considered 'sufficient' or when references 'fit' the inquiry question. Hence, with such descriptions it is difficult for novices to determine the quality of their own work, let alone to improve it. Another problem is that in general rubrics are not tested for interreliability (by different researchers). The same applies for validity: does the instrument suit the items that should be evaluated? (Ledford & Sleeman, 2000). Moreover, Jonsson and Svingby (2007) showed that rubrics mostly focus on the assessment of quality in student essays and reports and not on the evaluation of inquiry processes.

In a previous study, the feasibility of a set of rubrics based on four design characteristics was tested in class (see Chapter 3). The outcomes of this study led to a refinement in the design characteristics regarding the number of rubrics and the need for an extension of the self-evaluation instrument with two extra tools. The aim of the present study was to gain insight into the feasibility of the revised and extended self-evaluation instrument when pre-university science students use it to evaluate ARV in inquiries in different school science subjects.

4.2 THEORETICAL FRAMEWORK

The refined design characteristics of the self-evaluation instrument (hereafter called the 'Evaluation of Quality of Inquiries (EQI) instrument') with which students can evaluate the ARV in inquiries were based on the outcomes of the first test cycle (see Chapter 3). The design characteristics of the EQI instrument, described further on in more detail, were related to:

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

saskia_proefschrift_e_chapter_04.indd 79 27-10-16 10:49

4.2.1 Content

Regarding the first design characteristic, content of the EQI instrument, recent studies on learning scientific inquiry in the school subjects of biology and chemistry (Schalk, Van der Schee, & Boersma, 2009; Van Rens, Pilot, & Van der Schee, 2010) showed that the concepts of evidence (CoE) model (Gott, Duggan, Roberts, & Hussain, n.d.) contains concepts that are feasible for increasing students' procedural knowledge. Gott et al. (n.d.) described 82 concepts of evidence (CoE) that can be important in building up evidence in a scientific inquiry. This includes evaluating the ARV of inquiries in those school subjects. Based on the explorative study (see Chapter 2), 23 CoE were selected to be elaborated in a set of 19 rubrics with which the ARV of inquiries can be evaluated. These CoE were rewritten for use by pre-university students and are hereafter called (CoE) items. From the first research cycle (see Chapter 3), it was concluded that 21 of the 23 items are relevant for pre-university students as novices when they evaluate the ARV in inquiries. From these 21 CoE items, 13 items are crucial for the students both when they evaluate ARV in their inquiry plans and in completed inquiries, therefore these 13 should to be worked out in rubrics. It was also concluded that students need two extra tools: an overview tool or ARV card and an easy-to-use checklist to handle all 21 items. At the level of the student novices, related to actions like checking whether a measuring instrument has been reset before the next measurement, it seems to be appropriate to include these kind of items only in an easy-to-use list so that the students can check whether they have performed all necessary actions in the various phases of an inquiry. The holistic or ARV card contains a short description of all 21 CoE items and their relation to the ARV of the various phases in an inquiry. Table 4.1 presents the 21 items from the CoE model with the items in the EQI instrument that were incorporated in a rubric.

4.2.2 Extent of complexity

The items as elaborated in the checklist and ARV card can be considered as 'done' or 'not done' and do not need any complexity in their descriptions of levels of performance. Therefore, the second design characteristic, the extent of complexity of the description of the EQI instrument, is related only to the 13 CoE items that need to be worked out in rubrics. According to Jonsson and Svingby (2007), the descriptions in rubrics need to have a hierarchical order in which the different levels of complexity of an (CoE) item become visible. Consequently, the students and teachers have to determine at which level of complexity the related part of the inquiry is performed and how the students can (learn to) perform at a more complex level.

The Structure of Observed Learning Outcomes (SOLO) taxonomy seemed to be applicable for the hierarchical descriptions in the rubrics (Biggs & Tang, 2007). Chan, Tsui, Chan, and Hong (2002) concluded that this taxonomy is the most suitable for determining various kinds of student learning outcomes. Moreover, the SOLO taxo-

saskia_proefschrift_e_chapter_04.indd 80 27-10-16 10:49

nomy supported undergraduate students when evaluating their activities at different points during the enactment of a learning task (e.g. Hodges & Harvey, 2003; Levins & Pegg, 1993; Minogue & Jones, 2009).

Table 4.1
Twenty-one items from the CoE model in the EQI instrument as incorporated in a rubric. All items were incorporated in the checklist and the ARV card

Item from CoE model	**Rubric**
Logical reasoning in theoretical framework	Theoretical framework[a]
Same (in)dependent variables in whole inquiry	
Specific and concrete inquiry question	Inquiry question
Hypothesis can be tested by inquiry method	Hypothesis
Inquiry method fits inquiry question	Inquiry method
- Sample is large enough	Drawing a sample
- Sample is sufficiently varied	
Measure and observe with more than one observer	
Measuring instrument is calibrated	
Reduce influence of (other) variables	
Conduct control experiment	
Repeat measurements, calculate average and deviation	Average and deviation in measurements
Objectivity in measuring and observing	
Systematic measuring and observing	
Drawing conclusions	Answer the inquiry question
Use of evidence[b]	The evidence in the conclusion
Measuring instrument is sufficiently accurate	Evaluation of accuracy
Sufficient results to infer conclusion	Evaluation of reliability
Evaluation of validity	Evaluation of validity
Comparability of results with other inquiries	
Recommendations for supplementary inquiries	Suggestions for future inquiries

a *This rubric was not used in the studies as described in Chapters 4–6. In the successive inquiry units, the theoretical framework was given to the students. Therefore, it was not necessary or possible to evaluate the validity of a by the students' written theoretical framework.*
b *When designing the revised set of rubrics, it was decided to make two rubrics out of the rubric 'Drawing conclusion and use of evidence' (used in the first research cycle) in order to let the content better fit the performances of the students.*

The SOLO taxonomy distinguishes five performance levels of increasing complexity: prestructural, unistructural, multistructural, relational and extended abstract. In evaluating the ARV of an inquiry, the prestructural level is visible when students make use of daily language, for example, 'we did it as precisely as possible', to describe accuracy. When students use one aspect, mostly by imitation of the teacher's language, then the level of performance is considered as unistructural. At the multi-

saskia_proefschrift_e_chapter_04.indd 81 27-10-16 10:49

structural level of performance, students describe more aspects that are relevant without considering inconsistencies or possible relations between aspects. At the relational level of performance, students consider inconsistencies or relations between aspects. Furthermore, when students can indicate how their inquiry results are related to inquiries in the respective domain, they have attained the extended abstract level. If the SOLO taxonomy has been worked out correctly in a rubric, then it is expected that a student will only be able to perform at a certain level when he or she has attained all the preceding levels of complexity (Biggs & Tang, 2007). Hence, a performance can only be evaluated at the relational level when this performance also fulfils the described performance at the multistructural level.

4.2.3 Extent of detail

The third design characteristic involved the extent of detail that is included in the rubrics. Rubrics can represent levels of performance in a holistic or an analytic way. An analytic instrument seemed to be the most appropriate for self-evaluation by novices (Arter & McTighe, 2001; Mertler, 2001). Mertler (2001) expected students to be able to focus attentively on their work with an instrument with an analytical (step-by-step) character and be informed by the instrument on how to improve their performance. In particular, students who are less experienced in self-evaluation learn more from using rubrics with an analytic character than from those with a holistic character. However, in a previous study (see Chapter 3) it was observed that students see the single rubrics as isolated tools to evaluate the ARV in various parts of an inquiry and not as part of a set of rubrics that can be used to evaluate the ARV of the complete inquiry. To create more coherence between the rubrics, it was decided to design a holistic overview tool (ARV card) as part of the EQI instrument (see 4.2.1).

4.2.4 Extent of general application

The fourth design characteristic concerned the extent of general application of the rubrics. Rubrics can have a specific function, for example, for 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). Because our goal was for students to evaluate inquiry tasks in different school science subjects with the same (EQI) instrument, generic rubrics were considered more applicable than task-specific ones. A point of concern in the use of generic instruments, in particular for novices, was transfer; the students gain knowledge in a certain context and they need to apply it in other contexts (Bransford, 2000). This transfer could be supported, as Jonsson and Svingby (2007) argued, by adding benchmark samples for each of the level descriptions in a rubric. These benchmark samples in rubrics were expected to help students to interpret the generic descriptions in the rubrics as is intended.

saskia_proefschrift_e_chapter_04.indd 82 27-10-16 10:49

Jonsson and Svingby (2007) suggested making the examples as diverse as possible and relating them to different contexts. However, as concluded in Chapter 3, too great a diversity in the examples of the different SOLO taxonomy levels in the rubrics confused students. They misunderstood the hierarchy in the levels in a rubric and the coherence between performance levels. Hence, the norm-referenced examples in the rubrics needed to be adapted to one inquiry theme that was general enough for students to switch easily to inquiries in different school science subjects and that was broad enough to be worked out in all levels of the eleven rubrics. Table 4.2 presents an overview of the four refined design characteristics for the EQI instrument with which pre-university science students could evaluate the ARV in inquiries.

Table 4.2
An overview of the four design characteristics for the revised EQI instrument with which pre-university science students can evaluate the ARV in inquiries

Design characteristics of the revised EQI instrument
1. The *content* of the instrument is based on 21 relevant items from the CoE model. Twelve CoE items are described in detail in eleven rubrics and all 21 items are described as actions in a checklist and an overview tool (ARV card).
2. The *extent of complexity* of the descriptions in the instrument concerns the eleven rubrics. Each rubric contains descriptions in five hierarchical levels of complexity that are based on the SOLO taxonomy. The descriptions are attuned to the level of pre-university science students.
3. The *extent of detail* in view of the use by novices led to rubrics with an analytical, step-by-step character, supplemented with an overview tool to show the coherence between the eleven rubrics.
4. The *extent of general application* in inquiries in different school science subjects requires transfer to occur. Each of the eleven rubrics contains norm-referenced examples regarding one inquiry topic to elucidate the general description in the hierarchical levels of complexity.

4.3 RESEARCH QUESTION

Before the learning outcomes from using the EQI instrument could be determined, it was important to know whether the refined design characteristics contributed to the feasibility of the instrument for the self-evaluation of ARV by pre-university science students in an inquiry. When the instrument had insufficient feasibility this had to be improved before the student learning outcomes could be studied. The research question in this study was:

> *What is the feasibility of the EQI instrument for the evaluation of the accuracy, reliability and validity in inquiries in different school science subjects by pre-university science students?*

saskia_proefschrift_e_chapter_04.indd 83 27-10-16 10:49

4.3.1 Determining feasibility

Regarding the feasibility of the EQI instrument, it needs to perform three functions in the teaching-learning process: (1) it can be used for self-evaluation of the ARV by students in an inquiry, (2) it can give information to the teacher on what support the students need when evaluating the ARV in an inquiry, and (3) it can develop an 'inquiry language' with which students and teacher can communicate on the ARV in an inquiry. Based on these functions, four criteria were derived with which a reflection could be made as to whether the design characteristics lead to a self-evaluation instrument that shows feasibility in supporting the intended teaching-learning process (Nieveen, 2009). These criteria were:

a. the EQI instrument is used as intended;
b. students and the teacher can work with the EQI instrument;
c. the EQI instrument supports the determination of the level of student understanding;
d. the EQI instrument leads to a positive change in use of 'scientific' terminology by students and teachers.

Table 4.3 shows the criteria to reflect on the feasibility of the EQI instrument related to the four design characteristics.

Table 4.3
Criteria to reflect on the feasibility of the EQI instrument related to the four design characteristics

Criteria to reflect on feasibility of EQI instrument	Design characteristics			
	1	2	3	4
a. EQI is used as intended		√	√	
b. Students and teacher can work with the EQI	√	√	√	
c. EQI supports determination of level of student understanding		√	√	√
d. EQI leads to positive change of 'scientific' terminology	√		√	√

4.4 METHOD

The feasibility of the EQI instrument was determined by a formative evaluation as part of a design-based research (Van den Akker, Gravemeijer, McKenney, & Nieveen, 2006). A qualitative method (Cohen & Manion, 1994) with triangulation of data (Yin, 2003) was used.

4.4.1 Description of the EQI instrument

The EQI instrument as designed contained an ARV card and a checklist as well as twelve rubrics. For the complete student EQI instrument, see Appendix C.

The rubrics were meant for students to evaluate the ARV in their inquiry plans as well as for reflection on a completed inquiry. Twelve items from the CoE model were

saskia_proefschrift_e_chapter_04.indd 84 27-10-16 10:49

worked out into twelve rubrics. Each rubric contained five levels of increasing complexity with norm-referenced examples. The examples in all rubrics and at all levels are related to the inquiry theme 'Measuring the human body when exercising'. When conducting their evaluations, the students circled, for each rubric, the level that they thought they had achieved. The rubric 'Theoretical framework' was not used in the teaching-learning process, because it was decided to provide the students with a theoretical framework instead of letting them write this framework themselves (as was done in the first test cycle). Hence, eleven rubrics are relevant for this study.

With the checklist, the students checked whether they had done all the necessary steps in their inquiry. The checklist (see Appendix C) posed questions regarding ARV on the preparation of the inquiry and on the execution of the inquiry, for example: 'Did you indicate in your inquiry plan how many times you want to repeat each measurement?' These questions were derived from the CoE items that were selected from the CoE model (Gott et al., n.d.). By using the checklist, for each of the questions the students filled out one of the categories: 'done', 'partly done', 'not done', or 'not relevant in this inquiry'. When students filled out the categories 'partly done' or 'not done', they saw an indication on the checklist on how to improve the unfulfilled item. The ARV card (see Appendix C) contained a short description of all 21 CoE items and their relation to the ARV of an inquiry. The students used this tool as an overview document, for instance, when they made a general judgement on the ARV of a completed inquiry.

Preceding this study, the EQI instrument was independently analysed by two researchers for its suitability and completeness for evaluating the ARV in inquiries, for the usefulness of the learning materials in the different inquiry units and for their usefulness in various phases during an inquiry. These analyses of both researchers showed an agreement of 89%, so it was concluded that the EQI instrument was suitable to be used in the study in a classroom setting.

4.4.2 Intended use of the EQI instrument

The students used the EQI instrument in three successive inquiry units (see http://hdl.handle.net/1871/38422). The inquiry units were in general science, biology and physics, respectively. In the first inquiry unit, the EQI instrument was introduced by evaluating the ARV of a research study on probiotics. Next, the students conducted their own inquiries about the cooling rate of hot coffee; about taste on the human tongue (in the biology inquiry unit); and finally, in the physics inquiry unit, about improving the traffic situation at a dangerous crossing by measuring the speed and braking distances of cyclists. From a pilot study, it was concluded that the inquiry tasks in the units were open and stimulating enough for the students to design and conduct their own inquiries (see Appendix D). In Figure 4.1 an overview is given of the three successive inquiry units.

saskia_proefschrift_e_chapter_04.indd 85 27-10-16 10:49

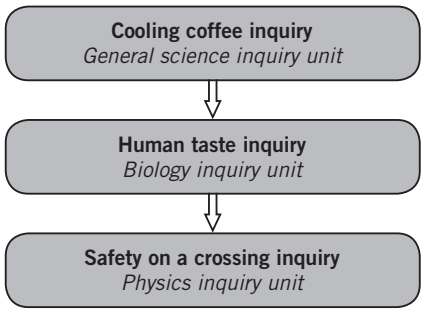

Figure 4.1
Overview of the successive inquiries in the three inquiry units

In each inquiry unit, the students studied a theoretical exposé. Based on the problem in the exposé each of the groups prepared its own individual inquiry plan with an inquiry question, a hypothesis and a method for how to answer the inquiry question. For example, in the biology inquiry unit, the theoretical exposé concerned human taste on the tongue with two experiments that had contradictory results and also led to contradictory conclusions and theories. Consequently, the students designed an inquiry plan in order to find evidence to support one of the contradicting theories.

With the checklist for inquiry preparation at hand, the students checked their inquiry plans and changed them if they thought that it was necessary. After this, the students evaluated their plans using the relevant rubrics in the EQI instrument. Then they could make the final changes to their plans. In the next step, they conducted the inquiry set-up they had designed and checked their performances with the checklist for conduct of the inquiry. This was followed by analysing the results and using them as evidence to support conclusions. Finally, they evaluated the ARV of their completed inquiry using the relevant rubrics in the EQI instrument. They could use the ARV card at all times during the inquiries. After each inquiry unit the students were explicitly asked to note any additions and clarifications on their ARV cards.

The teacher also checked and evaluated all the students' inquiry plans and completed inquiries with the same checklist and rubrics as in the student EQI instrument. The teacher's checks and evaluations were handed to the students at the start of each lesson. Based on a comparison of their checks and evaluations with those of the teacher, the students could discuss contradictions with the teacher and decide to make adaptations.

For each of the inquiry units a workbook was designed with activities for the students. During all lessons, each student had the complete EQI instrument at hand. The intended use of the instrument was explained to all students before they started the first inquiry unit. During the first inquiry unit, they first applied the instrument to

saskia_proefschrift_e_chapter_04.indd 86 27-10-16 10:49

an existing piece of scientific research on probiotics before they conducted their own inquiries on the cooling rate of coffee. The terminology in the instrument was elucidated in the workbook for each inquiry unit and explained by the teacher at the start of each unit. The workbooks contained indications for the student on at what points in the teaching-learning process they needed to use the checklist, the rubrics or the ARV card. As a result of their checks and evaluations, the students were stimulated to improve their work. All inquiry units were also guided with a teacher's manual in which the intended use of the student EQI instrument as well as the guides for instruction to enact the unit were described (see http://hdl.handle.net/1871/38422).

4.4.3 Participants

Twenty-seven pre-university science students, aged 16 or 17, voluntarily participated in the second test cycle. They signed up to participate after receiving an e-mail about the study from their science teachers. All the students were studying the school science subjects of 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 which was conducted after regular school periods.

4.4.4 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 composition of the groups was the same in all units. The students' work was done in seven afternoon sessions, each of three hours, over a time span of three months. The decision to conduct this study outside of 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 in this way of organising this research.

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 the teacher of the various inquiry units in the second research cycle. The researcher was not the students' normal biology teacher, but stood in as teacher to perform this research study. Being one of the designers avoided misinterpretation by the teacher of both the content of the instrument and the intended teaching-learning process, as happened during the first research cycle. 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. Two precautions were taken in the research procedure to minimise the effect of the dual role of teacher and researcher that could be a limitation in the study. First, two observers, who were not involved in the design process of the EQI instrument, were present in all ses-

saskia_proefschrift_e_chapter_04.indd 87 27-10-16 10:49

sions to video-record and observe whether the lessons were taught as designed and whether the teacher influenced the data collection. Second, all data were independently analysed by two researchers. When any differences occurred, a discussion took place until consensus was reached.

4.4.5 Data collection and data analysis

The two observers made a field report of their observations after each session. Moreover, all sessions were video-recorded. The discussions of all student groups and the communications of the teacher were audio-recorded and transcribed. After each session the teacher and the observers sat together to talk about the use of the EQI instrument by the twelve student groups; these gatherings were also audio-recorded and transcribed. Immediately after this, the teacher wrote a reflection report in which she compared the intended use of the instrument with the actual use of it in class. The observers read each reflection report and, when any difference occurred, adaptations were discussed until consensus was reached.

Immediately after each session, the students individually completed questionnaires on the feasibility of the instrument. Over the seven afternoon sessions, 187 questionnaires[4] were collected. One week after the last session, all the students were interviewed in pairs on their experiences with the feasibility of the instrument. These pairs were not necessarily the same as the student groups during the afternoon sessions. The interviewer had not been present in the sessions. All interviews were transcribed.

Moreover, the completed student and teacher checklists, rubrics and ARV cards in the successive inquiry units, as well as all student group inquiry plans and workbooks with their notes, were collected. All completed student group checklists, rubrics and ARV cards in the successive inquiry units, all field reports, all teacher reflection reports and student workbooks were analysed to determine the actual use made by students of the instrument in class. After this, a comparison was made in order to determine whether the students had used the EQI instrument as was intended (criterion a). This criterion would be met when at least 80% (≥ 22 students) used the instrument as was intended (c.f. Juran, Gryna, & Bingham, 1974).

To determine whether the students could work with the EQI instrument in different inquiry phases and in different school science subjects (criterion b) all the students' responses in the questionnaires and interviews were analysed. At least 22 students (>80%) should respond positively for the instrument to be judged a manageable instrument (Juran et al., 1974). The teacher's reflective report and the observers' field reports were analysed for the feasibility of the instrument for the teacher.

4 Two students were absent during one of the afternoon sessions and therefore did not fill out a questionnaire.

saskia_proefschrift_e_chapter_04.indd 88 27-10-16 10:49

Regarding criterion c, handling the EQI instrument supports the teacher in determining what help the students need when evaluating the ARV in inquiries, the student scores in all rubrics and checklists were compared to the teacher's scores. When the scores of at least six groups (50%) deviated from the teacher's score then extra support was needed in class (Hafner & Hafner, 2003). The field reports and the teacher's reflection reports were analysed for subjects that were discussed by the teacher and a student group.

To determine whether the use of the EQI instrument led to an increase in adequate 'scientific terminology' related to the evaluation of ARV inquiries (criterion d) the audio transcripts of the interactions among the students as well as between the students and the teacher in the general science unit (first unit) and the physics unit (last unit) were analysed. These transcripts were analysed for changes in students' use of items from the CoE model. Moreover, the same analysis was done on the students' evaluations of their inquiry plans and completed inquiries. In these analyses, a distinction is made between 'adequate' and 'inadequate' use of language (see Chapter 2). To achieve criterion d, at least 10 of the student groups (>80%) should show adequate language regarding one item on accuracy, one on reliability and one on validity. At least 80% of the items should be used adequately in the students' language in order to speak of adequate use of the CoE (Juran et al., 1974).

4.5 FINDINGS

4.5.1 Actual use of the EQI instrument (criterion a)

Analysis of the observers' field reports and the teacher's reflective reports showed that all twelve groups used the ARV card as was intended. Moreover, all twelve groups used the tool on one or more occasions during the inquiries. In their own inquiries in the three inquiry units eight groups of students regularly looked back to the activity in the first inquiry unit in which they studied the ARV card in another researcher's previous inquiry. Three students used the space for adding explanations on the ARV card.

Analysis of the checklist and rubrics showed that twelve groups of students completed these in each of the three inquiry units. Moreover, after use of the checklist at least ten student groups made changes to their inquiry plans regarding the inquiry problem in each of the three inquiry units. Furthermore, analysis of the observers' field reports, the teacher's reflective reports and the student workbooks showed that in each of the three inquiry units at least six groups used the descriptions and benchmark samples in the rubrics to reflect on and improve their inquiry plans and completed inquiries. The analysis also revealed that after completing the checklist no groups repeated their experiments.

saskia_proefschrift_e_chapter_04.indd 89 27-10-16 10:49

In summary, in each inquiry unit at least 80% of the student groups used the checklist, rubrics and ARV card of the EQI instrument to check and evaluate their inquiry plans and completed inquiries.

4.5.2 Students can work with the EQI (criterion b)

Analysis of the transcripts of the student group interviews revealed that 25 students (93%) experienced the checklist, rubrics and ARV card of the EQI instrument as useful in different inquiry phases in the three inquiry units. Analysis of the student responses in the questionnaires showed 156 responses from 27 students regarding the handling of the EQI instrument. The student responses are equally divided over the three inquiry units. Positive aspects of the EQI instrument appeared in 129 student responses (83%), for example:

- *I then realised how bad our inquiry question was.*
- *I could quickly see what went wrong as well as what I had forgotten.*

Thirteen student responses (8%) also showed positive aspects as well as limitations in handling the EQI instrument, for example:

- *It is good to look at your inquiry in such a way but I also need the support of the teacher, because to find everything out for yourself is difficult.*

Ten student responses (6%) revealed limitations in handling the instrument, like:

- *In the taste experiment, it was impossible to collect enough measurements to get a reliable inquiry and to evaluate that with the tables [rubrics].*

Four of the student responses (3%) could not be categorised.

Analysis of the teacher's reflective reports showed that the students completed the checklist and rubrics without difficulties. All CoE items that were worked out in the rubrics fitted at least two of the three inquiry units. All students could evaluate all phases of the inquiries in the three successive units. The teacher used the benchmark samples in the rubrics when she hesitated between two levels in a rubric. Moreover, analysis of the teacher's reflective reports revealed that, according to the teacher, the students' scores on a certain level in the rubrics demonstrated that they had mastered the previous level(s). Furthermore, the analysis showed that the teacher in four rubrics – Inquiry question, Evaluation of accuracy, Evaluation of reliability and Evaluation of validity – needed more levels than the five levels of the SOLO taxonomy to indicate better how far the students had progressed.

Analysis of the observers' field reports revealed that the teacher completed the checklist about performing the inquiry in the first two inquiry units for an average of seven student groups while in class. In her reflective report regarding this situation in class, she wrote:

saskia_proefschrift_e_chapter_04.indd 90 27-10-16 10:49

- *Because of limited time, I could only complete the checklist for some groups. The groups that I could not give written feedback got oral feedback. After this feedback, these groups still discussed possible changes.*

Further analysis of the field reports showed that, during the conduct of an experiment, eight groups of students calculated the average of measurements that were measured with different thermometers. Even after the students realised that the measurements showed great deviation, they themselves made no link to the accuracy of the measurements. At that point the teacher gave oral feedback to the eight groups and marked on the checklist that the list also needed a question about the use of different measuring instruments and the influence of that on the accuracy of the measurements.

Hence, more than 80% of the students as well as the teacher could handle the EQI instrument in different inquiry phases and in inquiries of different school science subjects. It was observed that the teacher had too little time to complete in class the checklist about performing the inquiry.

4.5.3 EQI supports determination of the level of student understanding (criterion c)

Analysis of the rubrics of the general science inquiry unit showed that, in seven of the eleven rubrics, six or more student groups (≥ 50%) had the same rubric score as their teacher. In the biology inquiry unit, five out of eleven rubrics were scored similar to the teacher's score by at least six student groups. In the physics inquiry unit, one out of the four rubrics was scored similarly to the teacher's score by at least six student groups (see Table 4.4). It should be noted that the students could not complete their inquiry on improving the traffic situation at a dangerous crossing because of very bad weather.

Table 4.4

Percentage of student groups with the same rubric scores as the teacher in the general science, biology and physics inquiry units

Rubric	General Science (%)	Biology (%)	Physics (%)
Inquiry question	50	50	33
Hypothesis	42	42	50
Inquiry method	50	25	25
Drawing a sample	58	17	33
Average and deviation of measurements	58	n.a.	-
Answer to the inquiry question	42	50	-
The evidence in the conclusion	25	50	-
Evaluation of accuracy	67	25	-
Evaluation of reliability	50	58	-
Evaluation of validity	50	75	-
Suggestions for further inquiries	42	17	-

Note. n.a.: not applicable; - : not conducted

saskia_proefschrift_e_chapter_04.indd 91 27-10-16 10:49

Another aspect of the analysis is the content of the discussions between the students and the teacher on the differences in their checklist and rubric scores. Analysis of the teacher's reflective reports showed that the teacher noticed that in each inquiry unit three (25%) or four (33%) student groups showed more knowledge about the CoE items during the discussion than is visible in the scores and in their workbook notes. Regarding this, the teacher wrote:

- *I ought to complete the rubrics after the students have explained their inquiry plan […] Maybe, a short presentation on their inquiry plan is a good idea.*

Analysis of the observers' field reports revealed that the students lacked sufficient knowledge on the items 'Sample is large enough and sufficiently varied' and 'Conduct control experiment'. When the teacher noticed this lack of understanding during the general science inquiry unit, she addressed these two items in a whole class conversation. After this explanation, the students showed better understanding of these two items. Regarding this, a student group wrote on their worksheets of the biology inquiry unit:

- *A conclusion that is based on five test subjects is not very reliable; the sample should be much larger to know something about the tasting of all adolescents.*

Further analysis of the field reports and the filled-out checklists showed that seven student groups asked their subjects in the taste on the tongue inquiry to rinse their mouths between the various experiments, which they considered as a reset of the measuring instrument. Students seemed to handle the CoE item 'Measuring apparatus is calibrated' in a different way to what was meant by the designers of the EQI instrument.

In summary, in various CoE items, 50% or more of the student groups showed a different score than the teacher's score. However, further analysis showed that those differences were frequently caused by the students' use of unclear terminology in their workbooks and EQI instrument.

4.5.4 Increase in 'scientific' terminology (criterion d)

For this criterion, the completed inquiries of the general science inquiry unit (first unit) and the physics inquiry unit (last unit) were compared as well as the transcripts of the audio recordings of the two lessons in which the students completed these inquiries. Analysis of the students' completed inquiries in the general science inquiry unit and their statements in the transcripts of the audio recordings showed that eleven of the twelve student groups mentioned at least one CoE item in their evaluation of the ARV. The eleven groups together wrote about 13 of the 20 items. Six student groups (50%) each wrote or mentioned at least about one item on accuracy, one on reliability and one on validity (all three categories). One student group wrote

saskia_proefschrift_e_chapter_04.indd 92 27-10-16 10:49

about items from two categories and four student groups wrote about items from one category. These student writings contained adequate language for the CoE items in 38% of the answers, like:

- *Our conclusion […] is based on the results, so the inquiry is valid* (drawing of conclusions).

Analysis of the students' evaluations of their inquiry plans in the physics inquiry unit revealed that the twelve groups each wrote or mentioned at least one item on accuracy, one on reliability and one on validity. At least one student group wrote about all 21 CoE items, the other eleven groups mentioned between 3 and 18 items. The student writings contained adequate language for the items in 79% of the answers, like:

- *We will take care of influencing factors on the speed of the cyclists like the direction and the strength of the wind* (control of (other) influencing variables).
- *We need an accurate speed-sense computer to measure the speed of our cyclists. Probably first test, then use* (accuracy of measuring instrument).

In summary, the student self-evaluations showed an increase from 50% to 100% in the use of at least one item on accuracy, one on reliability and one on validity. The use of adequate 'inquiry language' increased from 38% to 79%. The criterion was 80%, so this was nearly fulfilled.

4.6 DISCUSSION AND CONCLUSION

Four criteria needed to be fulfilled to determine the feasibility of the EQI instrument in evaluating the ARV in inquiries by pre-university science students. Criterion a, on *the students' use of the instrument as intended,* and criterion b, on *students and teacher can handle the EQI instrument* in different inquiry phases and different science subjects, were fulfilled completely. Criterion c on *the EQI instrument supports the determination of the level of student understanding* and criterion d on *the use of the EQI instrument leads to an increase in adequate 'inquiry language' related to the evaluation of ARV in inquiries* were sufficiently fulfilled.

As presented in the results, the three parts of the instrument – rubrics, checklist and ARV card – were adequately used by most of the students at the intended points in the three inquiry units, so criterion a *the EQI instrument is used as it was intended* is fulfilled. Second, criterion b, *students and the teacher can work with the EQI instrument* was also fulfilled. Most of the students responded in the interviews that they could adequately handle the instrument in an inquiry process as well as in inquiries in different school science subjects. This was in accordance with the teacher's view, as was visible in her reflective reports. From the result that criterion b was fulfilled it can be concluded that the descriptions of how to use the EQI instrument were attuned to the level of novices.

saskia_proefschrift_e_chapter_04.indd 93 27-10-16 10:49

In the study on the first design of the EQI instrument (see Chapter 3), it was concluded that it was too bulky. In the present study, no indications regarding the size of the instrument were found. However, the teacher faced a problem in completing all the checklists on the conduct of the experiment in time. It may be unnecessary for the teacher to complete those lists, because direct oral feedback from the teacher with the checklist at hand enabled students to make adjustments in the conduct of their inquiry plans. To be sure that the teacher can adequately use the CoE items from the checklist in his or her oral feedback, the teacher should be an expert with sufficient knowledge on these CoE items (Bransford, 2000). Most teachers are not used to supervising inquiries in which students focus on the evaluation of ARV. Sometimes they know the concepts ARV and the CoE from their academic science studies, but in teacher training courses little attention is paid to this aspect of learning to inquire. Due to the new standards on the evaluation of ARV in inquiries in the new formal science curriculum of biology, chemistry and physics in the Netherlands, it would be worthwhile to develop an in-service course for teachers to learn more about this specific focus on learning to inquire. The desire to have a teaching-learning process on learning to inquire in different school science subjects also motivates teachers to learn about the evaluation of ARV in inquiries. Consequently, two objectives should be included in in-service training: as well as learning a 'new language' to speak with the students, the teachers should learn how they could provide formative feedback on ARV in their students' inquiries, for example, with help of the EQI instrument (Van der Schee & Rijborz, 2003).

From the teacher's reflective reports, it could be concluded that the checklist lacked a question on the comparison of measuring instruments. This appeared to be a further elaboration of the CoE item 'Control of (other) influencing variables' which is part of the checklist and ARV card. Moreover, a further elaboration of the rubric 'Evaluation of reliability' was needed when students conducted an inquiry with few experimental test subjects.

Third, criterion c, *the EQI instrument supports the determination of the level of student understanding,* was sufficiently fulfilled. Based on the checklists and rubrics that the students completed, the teacher could determine whether student groups need support as well as whether an explanation to the whole class was needed for certain CoE items.

Comparison of the students' and teacher's rankings in the rubrics did not produce the expected value, because for a lot of the students their written products were not a good indication of their mastery of a CoE item due to their use of unclear terminology. From the results about the necessary support for the students, it could be concluded that during an inquiry, students sometimes had certain CoE items in mind but did

saskia_proefschrift_e_chapter_04.indd 94 27-10-16 10:49

not write them down. The use and understanding of these items were invisible to the teacher because she only used the written inquiry reports. The observation that students wrote less than what they knew could indicate a shift from instrumental to mental handling of the CoE items by the students, which implies that the students' understanding had increased and therefore they needed less support from the teacher (Bransford, 2000). Moreover, the differences between the students' and teacher's scores could be caused by a different interpretation of the descriptions in the rubrics and checklists between the students and the teacher, as found in the CoE item on calibrating the measuring apparatus. These differences in interpretation of an item can only become visible in oral discussions between students and teachers about differences in their checklist and rubric scores.

Finally, on criterion d, *use of the EQI instrument leads to an increase in adequate 'scientific' terminology related to the evaluation of ARV in inquiries;* during the last inquiry, the physics inquiry unit, all student groups adequately mentioned at least one item regarding accuracy, reliability as well as validity. Compared with the first, general science inquiry unit the students showed an increase in adequate inquiry language that almost fulfilled the set norm. Moreover, as part of evaluating the inquiry plans in the last inquiry unit the students commented on all CoE items. From these findings, it could be concluded that students showed a robust increase in using adequate inquiry language. The students moved from being novices towards becoming experts, shown by their level of communicating about ARV in inquiries by using 'inquiry language' as described in the EQI instrument (Bransford, 2000).

Regarding the fulfilment of the four criteria, it was concluded that the student EQI instrument sufficiently supported the intended teaching-learning process. Hence, it was considered plausible that the design characteristics supported the development of an instrument that helps students, as novices, in the process of evaluating the ARV in inquiries.

The items from the CoE model, design characteristic 1, appeared to be adequate for evaluating the ARV in an inquiry with the instrument. The elaboration of 13 items from the CoE model in rubrics, and all items from the model in a checklist and ARV card, led to an instrument that novices adequately applied in different inquiry phases and in different school science subjects. The 13 items that were elaborated in rubrics were, regarding their complexity, adequately worked out into five levels of complexity based on the SOLO taxonomy (design characteristic 2). The students understood the hierarchical structure in the rubrics with their description and benchmark samples. The addition of more sublevels, as was mentioned by the teacher during this study, led to a more detailed description of the extent of complexity and examples in which CoE items were elaborated for the student and teacher. This can probably induce a

saskia_proefschrift_e_chapter_04.indd 95 27-10-16 10:49

better accord between the students' and teacher's scores in the rubrics (Chan et al., 2002). However, too detailed descriptions can reduce the transfer of the content to inquiries of different school science subjects. Hindering the transfer is undesirable when, as in this research study, the students have to use the instrument in inquiries in different school science subjects. As shown in the analysis of criteria a and b, the ARV card was indeed used by the students as an aid to evaluate a completed inquiry to come to a general judgement regarding the ARV in inquiries (design characteristic 3). In using the ARV card, no problems were found during the data analysis.

Finally, the three parts of the EQI instrument and most of the CoE items were suitable for evaluating the ARV in the three different inquiries (design characteristic 4). All CoE items could be used in at least two of the three inquiry units. The students were able, by using all the items, to evaluate the ARV in inquiries as intended. During the data analysis, no problems with the benchmark samples were found, so these samples, all around the same inquiry topic, were supportive in using the rubrics.

In conclusion, the EQI instrument is feasible for pre-university science students to evaluate the ARV in inquiries regarding different school science subjects. What the 'ideal' teaching-learning process is, in which pre-university students adequately evaluate the ARV in inquiries, and whether the students benefit, was the focus of the next studies of the second research cycle. These two studies are described in Chapters 5 and 6 respectively.

saskia_proefschrift_e_chapter_04.indd 96 27-10-16 10:49

REFERENCES

Aarsen, M., & Van der Valk, T. (2008). Onderzoekende houding, een leerlijn [Investigative approach, a teaching-learning trajectory]. NVOX, 33(8), 354–356.

Abd-El-Khalick, F., Boujaoude, S., Duschl, R., Lederman, N. G., Mamlok-Naaman, R., & Hofstein, A. (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. 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.

Burke, K. (2006). From standards to rubrics in 6 steps. Tools for assessing student learning (Revised ed.). Thousand Oaks, CA: Corwin 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. Assessment & Evaluation in Higher Education, 27(6), 511–527.

Chinn, C. A., & Malhotra, B. A. (2002). Epistemologically authentic inquiry in schools. Science Education, 86(2), 175–218.

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

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

Gott, R., Duggan, S., Roberts, R., & Hussain, A. (n.d). Research into understanding scientific evidence. Retrieved on 21 October 2013 from http://www.dur.ac.uk/rosalyn.roberts/Evidence/cofev.htm

Hafner, J., & Hafner, P. (2003). Quantitative analysis of the rubric as an assessment tool. 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, 785–787.

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.) (pp. 216–219). New York: McGraw-Hill.

Ledford, B. R., & Sleeman, P. J. (2000). Instructional design. Greenwich, CT: Information Age Publishing.

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. In S. K. Abell & N. G. Lederman (Eds.), Handbook of research on science education (pp. 393–442). Mahwah, NJ: Lawrence Erlbaum.

Mertler, C. A. (2001). Designing scoring rubrics for your classroom. Practical Assessment, Research & Evaluation, 7(25). Retrieved on 21 October 2013 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.

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.

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.

saskia_proefschrift_e_chapter_04.indd 97 27-10-16 10:49

Schalk, H. H., Van der Schee, J. A., & Boersma, K. Th. (2009). The use of concepts of evidence by students in biology investigations. 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 (ERIDOB). Utrecht: Beta 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.

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 on 21 October 2013 from http://PAREonline.net/getvn.asp?v=9&n=2

Van den Akker, J., Gravemeijer, K., McKenney, S., & Nieveen, N. (2006). Educational design research. London: Routledge.

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.

Van Rens, L., Hermarij, P., Pilot, A., Beishuizen, J., Hofman, H., & Wal, M. (2014) Pre-university chemistry students in a mimicked scholarly peer review. International Journal of Science Education, 37(4), 2514–2533.

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

saskia_proefschrift_e_chapter_04.indd 98 27-10-16 10:49

saskia_proefschrift_e_chapter_04.indd 99 27-10-16 10:49