Development of a Student Self-Evaluation Instrument in Inquiries

Saskia van der Jagt[1], Lisette van Rens[1], Herman Schalk[1], Albert Pilot[2] and Jos Beishuizen[1]

[1]Department of Research and Theory in Education, VU University Amsterdam, The Netherlands

[2]Freudenthal Institute for Science and Mathematics Education, Utrecht University, The Netherlands

Author note

Prepared for the NARST 2012 conference, Indianapolis, USA, March 24-28, 2012. Work in progress—Please do not cite without explicit permission of the authors.

Saskia van der Jagt, Lisette van Rens, Herman Schalk, Jos Beishuizen, Department of Research and Theory in Education, Faculty of Psychology and Education, VU University Amsterdam, The Netherlands; Albert Pilot, Freudenthal Institute for Science and Mathematics Education, Faculty of Science, Utrecht University, The Netherlands.

The authors thank Ludo Kolfschoten for his contribution to the rating and data-analysis.

Correspondence concerning this paper should be addressed to Saskia van der Jagt, VU University, Faculty of Psychology and Education, Department of Research and Theory in Education, De Boelelaan 1105, 1085HV Amsterdam, The Netherlands. E-mail: s.a.w.vander.jagt@vu.nl

Abstract

This educational design study aims at operationalizing design characteristics that lead to a for pre-university science students feasible self-evaluation instrument to evaluate the accuracy, reliability and validity in successive science inquiry units. A self-evaluation instrument with nineteen rubrics was designed. This design was based on four characteristics that were identified from the literature, among which the Concepts of Evidence model and the SOLO taxonomy. To determine the feasibility of the instrument upper secondary school students (n=24) used the self-evaluation instrument in class in three successive – general science, biology and physics – inquiry units. Data were obtained from written documents, audio- and videotapes, questionnaires and interviews. It is concluded that the self-evaluation instrument with rubrics seems to have the potential in learning pre-university science students how to evaluate the accuracy, reliability and validity of an inquiry. The four design characteristics are essential, but a major revision regards that part of the students self-evaluation instrument needs to become holistic instead of analytic. In the discussion and implications recommendations for further research and development are given.

Keywords: self-evaluation instrument, pre-university science education, SOLO taxonomy, concepts of evidence, inquiry.

Introduction

At secondary schools learning to inquire is becoming a more important part of the science education curriculum during the last decades (Abd-El-Khalick et al., 2004). 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 improve standard practical skills. Third, as a part of improving their procedural understanding, students learn how to evaluate the accuracy, reliability and validity of inquiries they conduct (Gott & Duggan, 1995; Millar, 2010).

This third objective is important in showing pre-university science students the cognitive processes of scientists in authentic inquiry contexts (Chinn & Malhotra, 2002). For students it is difficult to understand what is meant by evaluating the accuracy, reliability and validity of inquiry, because they are novices in evaluating these aspects. Mostly, inquiry tasks in school science subjects are like 'cookbook recipes' in which students follow the instructions rather mechanically and without reflection on the performance of the inquiry. In these 'cookbook-tasks' evaluating the accuracy, reliability and validity of an inquiry does not come in focus and as a result it is complicated for pre-university science students to improve their procedural understanding on the accuracy, reliability and validity of an inquiry (Lunetta, Hofstein, & Clough, 2007; Millar, 2010).

Transfer of this part of the procedural understanding to different inquiry contexts is even more difficult for pre-university science students, despite the similarities in evaluating accuracy, reliability and validity (Roberts & Gott, 2002). Transfer can 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).

From previous research it is known that novices in a certain domain, as pre-university science students in evaluating the accuracy, reliability and validity of an inquiry, should be provided with learning experiences in which they can recognize patterns in the domain and are supported in organizing new information and its connection to their prerequisite knowledge (Bransford, 2000). In organizing new information or knowledge for novices a self-evaluation instrument can have a useful supportive function. Andrade & Valtcheva (2009) showed that self-evaluation involves reflection on a task and revision of the work by students. The students said that self-evaluation helped them to focus on the main aspects of a task and to recognize the strength and limitations of their work.

A possibility for teaching pre-university science students how to evaluate the accuracy, reliability and validity of an inquiry might be to provide them with a self-evaluation instrument during an inquiry. Based on the research of Sevian & Gonsalves (2008) and from previous experiences in class of one of the authors we knew that a coherent set of rubrics could function as a self-evaluation instrument for pre-university science students during inquiries. Rubrics support learning by making performance criteria explicit, which makes it easier to give feedback to students and to let them perform a self-evaluation of their work. (Jonsson & Svingby, 2007).

These rubrics can be used as a formative instrument with qualitative descriptions of (levels of) performance criteria. However, many rubrics for secondary and higher education contain ambiguous descriptions of performance levels on skills and strategies across their scale levels and in general they are not tested on reliability and validity (Tierney & Simon, 2004). The review study of Jonsson & Svingby (2007) shows that most rubrics focus on the assessment of the content of student products (essays, reports) rather than on processes or strategies of students. More particularly, it is not known which characteristics of rubrics can help to improve the strategies of students in ensuring the accuracy, reliability and validity during the enactment of inquiries when they use the rubrics for self-evaluation. Therefore, our main research question is: Which design characteristics are needed to design a self-evaluation instrument that is feasible for pre-university science students to evaluate the accuracy, reliability and validity in successive science inquiry units?

Theoretical perspective

Design characteristics of a self-evaluation instrument with rubrics

In literature, we identified four design characteristics which seem to be useful for the design of the self-evaluation instrument for the aim of our study.

The first design characteristic is about the so-called ‘trait’ of the self-evaluation instrument or set of rubrics. For rubrics, this trait is mostly denominated as holistic or analytic (Arter & McTighe, 2001; Mertler, 2001). Holistic rubrics are seen as a means to make an overall judgment about the quality of a task whereas analytic rubrics are considered as a means to evaluate different, smaller aspects in a task. Analytic rubrics are also useful in giving specific feedback to students and for self-evaluating purposes (Arter & McTighe, 2001). Especially students with less experience in performing a specific task, in our case evaluating the accuracy, reliability and validity of an inquiry, learn more from using rubrics with an analytical trait than from using holistic rubrics. Therefore, for the aim of our study, we opt for a set of analytic rubrics, by which students learn in detail how to evaluate the accuracy, reliability and validity in different stages of that inquiry *(design characteristic 1*).

Depending on the application, a rubric can be specific for components of a single inquiry task (‘task-specific’) or can be used to evaluate the same components in various inquiry tasks (‘generic rubrics’). Generic rubrics can be used across analogous tasks, e.g. all inquiry tasks in school science subjects (Arter & McTighe, 2001; Jonsson & Svingby, 2007). Because of our goal to let students evaluate these aspects in inquiry tasks in different school science subjects, generic rubrics seem to be more applicable than task-specific ones *(design characteristic 2).*

This implicates a general description of the levels of performance of each rubric. To elucidate these general descriptions, we decided to provide the students with a benchmark sample for each of the descriptions. As Jonsson and Svingby (2007) argued, benchmark samples in rubrics help the students to interpret the descriptions in the rubrics in a similar way as the teacher. It should be kept in mind that benchmarks should be chosen with a variety as wide as the tasks the rubrics are used for. For the rubrics of our study we have to select benchmark samples that are feasible for all school science inquiry units where the rubrics will be used.

Recent educational research in chemistry (Van Rens, Pilot, & Van der Schee, 2010) and biology (Schalk, Van der Schee, & Boersma, 2009) has shown that the use of concepts of evidence (CoE)-model (Gott, Duggan, Roberts, & Hussain, n.d.) can improve students’ procedural understanding, among which the ensuring of the accuracy, reliability and validity of an inquiry (Gott & Duggan, 2003). This suggests that the content of a self-evaluation instrument regarding accuracy, reliability and validity of an inquiry can be related to the items in the CoE-model that are connected to the accuracy, reliability and validity of the inquiry design, the actual measurments, the obtained data and the reasoning with evidence (Gott, Duggan, Roberts, & Hussain, n.d.) *(design characteristic 3).*

For the aim of our study we made a selection of nineteen items out of the CoE-model that are expected to be appropriate for evaluating the accuracy, reliability and validity of an inquiry by pre-university science students. These nineteen items were used in constructing a student self-evaluation instrument composed of nineteen rubrics. Table 1 presents the subjects and intended use of the rubrics during an inquiry.

Table 1. Overview of subjects and intended use of the rubrics in the self-evaluation instrument.

Intended use Number and subject	**Intend to evaluate**
After preparing the inquiry	
1. Theoretical framework	validity
2. Inquiry question	validity
3. Hypothesis	validity
4. Research method of an experiment or observation	reliability
5. Taking of a sample	reliability
6. Preparation of tables to note down data	validity
7. Preparation of handling & analysis of data	validity
After collecting the data	
8. Experiment: independent variable	validity
9. Experiment: dependent variable	reliability
10. Performing observations	accuracy
11. Mean & spread of measurement values	accuracy
After handling the data	
12. Handling of outliers in measurement values	accuracy
13. Comparability of results	reliability
14. Drawing conclusion & use of evidence	validity
15. Defining of patterns in results	validity
After evaluation of the inquiry	
16. Evaluation of accuracy of the measurements	validity
17. Evaluation of reliability of the results	validity
18. Evaluation of validity of the conclusion	validity
19. Recommendation for supplementary inquiries	validity

Each of the nineteen rubrics of the self-evaluation instrument needs to be described in performance levels so that the student can get a good orientation on the evaluation process (Jonsson & Svingby, 2007). 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 as ‘you are almost performing as described on the expert level’. These statements hardly give students any insight on how to improve their performance (Jonsson & Svingby, 2007; Mertler, 2001; Moskal, 2000). Furthermore, to have a promotional effect on student learning in evaluating the accuracy, reliability and validity of an inquiry, the descriptions of the different levels of performance and the benchmark samples in each rubric should be easy to be distinguished for students. To show the successive steps in evaluating the accuracy, reliability and validity of an inquiry, all descriptions and benchmark samples in the rubrics should be represented hierarchical (Arter & McTighe, 2001; Moskal, 2000).

Therefore, we needed a taxonomy that was intended to be useful in describing the levels of performance in a more sophisticated and hierarchical way. The Structure of Observed Learning Outcomes (SOLO) taxonomy uses five levels: prestructural, unistructural, multistructural, relational and extended abstract. This taxonomy was considered to be suitable for our study, because it focuses on the levels of learning outcomes and is supportive in evaluating students’ performance at a particular moment in a learning task (e.g. Biggs & Tang, 2007; Hodges & Harvey, 2003; Lake, 1999; Levins & Pegg, 1993; Minogue & Jones, 2009)

Table 2 shows, by using the example of an inquiry question, what the characteristics of the five levels of the SOLO taxonomy are. The prestructural and unistructural levels are supposed to be based on the prerequisite knowledge of the students, whereas in the self-evaluation instrument for our study the prestructural level has a link with the prerequisite knowledge about the meaning of accuracy, reliability and validity in everyday language or the ‘daily-life-context’. The unistructural level starts from the prerequisite knowledge in inquiries about the CoE-subject that will be described in a rubric. Based on the ideas behind the SOLO taxonomy, the multistructural, relational and extended abstract levels of a rubric should be hierarchical built on the unistructural level.

When this taxonomy is properly applied to the content of the self-evaluation instrument, one can only reach the relational level when the multistructural level is met completely. The prerequisite knowledge of students from daily life and about the CoE-subject and the potential execution of the three highest levels of the rubrics were explored in a previous study (Van der Jagt, Schalk, & Van Rens, 2011) *(design characteristic 4).*

The four described design characteristics cover the main characteristics of a rubric: its trait, the degree of generality, the content and the descriptions of levels of performance (Arter & McTighe, 2001). These four design characteristics in designing this self-evaluation instrument with rubrics, will provide feasible instrument that students can use to evaluate the accuracy, reliability and validity in an inquiry can be designed. For an overview of the four design characteristics see Table 3. An example of a rubric can be found in Appendix I.

Table 2. Examples of the five levels of the SOLO Taxonomy transcribed to the inquiry question

Level	**Transcribed to inquiry question**
Prestructural	The inquiry question is based on everyday knowledge
Unistructural	The inquiry question shows one variable
Multistructural	The inquiry question shows both the independent and dependent variables
Relational	The inquiry question shows both the independent and dependent variables and is related to relevant domain specific knowledge
Extended abstract	The inquiry question shows both the independent and dependent variables and is indicative regarding the extension of relevant domain specific knowledge

Table 3. Overview of design characteristics in the student self-evaluation instrument with rubrics

Design characteristics
1. The instrument has an analytic trait.
2. The instrument is generic for evaluating the accuracy, reliability and validity in inquiry units in the different school science subjects. It contains benchmark samples to elucidate the generic descriptions.
3. The instrument contains rubrics that are based on nineteen CoE: twelve rubrics for evaluating the validity, four for the reliability and three for the accuracy of an inquiry.
4. Each rubric has five hierarchical levels conform the SOLO taxonomy. Pre- and unistructural level are based on the prerequisite knowledge of pre-university students. Multistructural, relational and extended abstract levels are built in a hierarchical way on the unistructural level.

General criteria for designing an instrument

In combination with the four design characteristics that support the aim of the self-evaluation instrument for our study we also involved some general criteria for designing an instrument. For every instrument it matters that one needs to be sure that the users understand the content and intended use of the instrument. The designer of the instrument has to be sure that the instrument is understandable and attractive for the user, has precise language - to avoid misunderstandings of the content - and has consistent terminology. In addition, the instrument has to be valid and reliable. An instrument is valid when it tests or measures what is supposed to be tested or measured. Reliability means that when an instrument is used multiple times for the same purpose or with different raters, it leads to similar outcomes (Ledford & Sleeman, 2000).

In our study, we tested the draft version of the self-evaluation instrument before introducing the rubrics in classroom with teachers and teacher-students on the construct-related validity, the construct-related reliability and used language (Arter & McTighe, 2001; Moskal & Leydens, 2000; Tierney & Simon, 2004). In addition, two pre-university science students were asked to comment the used language to make sure that their peers could understand the meaning of a description in the instrument as intended. The content of the draft rubrics was revised whenever an inconsistency in the construct-related validity, construct-related reliability or language appeared during the test.

Methodology

To evaluate the feasibility of the designed instrument, a qualitative research method was used (Cohen & Manion, 1994) with a triangulation of data (Yin, 2003). This method was chosen because we wanted to test the feasibility of the self-evaluation instrument in a naturally occurring setting of students in class (Collins, Joseph, & Bielaczyc, 2004).

Participants

The participants in the study were 24 pre-university science students (age 16-17) from an upper secondary school in The Netherlands. In pairs the students conducted three successive inquiry units in general science, biology and physics wherein the self-evaluation instrument was implemented. All participating students were studying biology, physics and chemistry at the pre-university school level. In their science classes the students were used to do practical work, but they had not yet experiences in evaluating the accuracy, reliability and validity of an inquiry.

The three teachers who were involved in the study were all qualified and experienced upper secondary teachers: one in biology, one in chemistry and one in physics. To enable them to instruct the students to apply the self-evaluation instrument all three teachers followed a workshop with one of the researchers.

Data collection and analysis

In each of the three inquiry units every student group was asked to evaluate the accuracy, reliability and validity of their inquiries with rubrics. As can be seen from Table 1, some rubrics were used after writing the inquiry plan, others after collecting and handling the data and a subset after formulating the conclusion and discussion. The following data were collected:

- The inquiry plans, data sets, conclusions and discussions of the student groups.
- The rankings of the student in the rubrics from three inquiry units.
- Videotapes of the teachers' instructions on the use of the rubrics.
- Questionnaires about the use of the instrument immediately after the students had completed each of the three inquiry units.
- The opinions of four students on the use of the instrument, who were interviewed after completion of all three inquiry units.
- The opinions of the three teachers, who were interviewed directly after each lesson, on the use and feasibility of the instrument during that particular lesson.

The accuracy, reliability and validity of student group inquiry plans, data sets and conclusions and discussions were rated independently by two reseachers - with the same self-evaluation instrument as was used by the students - with an inter-rater reliability of 73%. Next, the researchers compared these researchers' reference ratings to the rankings of the student groups when they used the self-evaluation instrument during the successive inquiry units.

This comparison was used to determine first which of the nineteen rubrics were actually used by the students. Second, to establish whether the instrument was indeed feasible to function as a generic self-evaluation instrument to evaluate the accuracy, reliability and validity in the students' inquiries in the successive inquiry units and whether the benchmark samples contributes to the generic character of the instrument. Third, to determine whether each of the nineteen rubrics in the instrument had suitable hierarchical levels, whereby is focused on determining whether in each rubric the pre- and unistructural level was in line with student pre-requisite knowledge and determining whether the multistructural, relational and extended abstract level were hierarchical built on the unistructural level.

All data were independently analyzed by two researchers and discussed until consensus was reached (Janesick, 2000).

Findings

The data derived from test phase in class was analyzed on the feasibility of the self-evaluation instrument with rubrics for pre-university students.

Actual use of the instrument

Analysis of the students' rubrics after writing the inquiry plan reveals that nine or more of the twelve student groups filled out the rubrics on the *theoretical framework* (nr. 1), *inquiry question* (nr. 2), *hypothesis* (nr. 3) and *research method of an experiment or observation* (nr. 4) in the successive inquiry units. The other rubrics were scarcely or never filled out by the students.

Moreover, analyses of the filled out student rubrics on 'after collecting the data' showed that all student pairs filled out these rubrics once in one of the inquiry units, but in other units, they did not complete these four rubrics at all. Some of the student pairs wrote comments under the concerning rubric(s) as "this rubric is not applicable to my inquiry" or "I can't remember exactly how I performed my observations".

Furthermore, analyses of the filled out student rubrics on 'after handling their data' reveals that nearly all student groups filled out the rubrics on *comparability of results* (nr. 13) and on *drawing the conclusion & use of evidence* (nr. 14). Now and then a student pair evaluated the *handling of outliers* (nr. 12). The rubric about *defining patterns in results* (nr. 15) is never used by the student pairs during the successive inquiry units.

Last, the analyses on the part 'after discussing the inquiry' reveals that in each inquiry unit nine or more student groups filled out the rubrics on *evaluation of accuracy* (nr. 16), *evaluation of reliability* (nr. 17) and *evaluation of validity* (nr. 18). Half of the student pairs filled out the rubric on *recommendations for supplementary inquiries* (nr. 19). For an overview of the actual use of the 19 rubrics by the students in the successive inquiry units see Table 4.

Analyses in the student responses in the questionnaire regarding the actual use of the rubrics reveals reponses like: "Half of the rubrics I could not use, because I had not done these things during my inquiry". One teacher made a similar remark during one of the interviews and stated that students first had to learn which steps they have to make during the performance of an inquiry before they can use the rubrics for self-evaluation. As she said: "As long as students don't know what spread is and how to determine the spread in their data set they won't make this step during an inquiry and can't evaluate their performance." Analyses of the observations in classroom and the video recordings support these comments.

Table 4. Actual use of the 19 rubrics by the student groups in the inquiry units.

Intended use during an inquiry Number and subject of a rubric	**Actual use by the students**
After preparing the inquiry	
1. Theoretical framework	+
2. Inquiry question	+
3. Hypothesis	+
4. Research method of an experiment or observation	+
5. Taking a sample	-
6. Preparation of tables to note down data	-
7. Preparation of handling & analysis of data	-
After collecting the data	
8. Experiment: independent variable	±
9. Experiment: dependent variable	±
10. Performing observations	±
11. Mean & spread of measurement values	±
After handling the data	
12. Handling of outliers in measurement values	-
13. Comparability of results	+
14. Drawing conclusion & use of evidence	+
15. Defining of patterns in results	-
After evaluation of the inquiry	
16. Evaluation of accuracy of the measurements	+
17. Evaluation of reliability of the results	+
18. Evaluation of validity of the conclusion	+
19. Recommendation for supplementary inquiries	±

Note. In the study twelve student groups participated. + stands for nine or more groups that used a rubric; ± for between five and eight groups; - for four or less groups.

Feasibility as a self-evaluation instrument in different inquiry contexts

The analysis also focused on the feasibility of the instrument for the function of a self-evaluation instrument for pre-university students in evaluating the accuracy, reliability and validity of students' inquiries in successive units. An indicator for the feasibility as a self-evaluation instrument is the agreement between the students' ranking in the rubrics and the researchers' rankings on the student group inquiry plans, data sets and conclusion and discussions since high agreement would indicate that students were able to use the instrument as intended by the designer. An instrument is considered to have the expected feasibility when 80% or more of the rankings of students and the researchers are similar (Juran, Gryna, & Bingham, 1974).

Analysis of the feasibility of the instrument showed 80% or more agreement in the ranking of the students and researchers in the *theoretical framework* (nr. 1), *inquiry question* (nr. 2), *taking a sample*[1] (nr. 5), *drawing conclusion & use of evidence* (nr. 14), *evaluation of reliability* (nr. 17) and *evaluation of validity* (nr. 18). Less than 40% agreement between the rankings was seen in the rubrics *research method of an experiment of observation* (nr. 4) and *evaluation of accuracy* (nr. 16). All rankings of students and researchers in the other rubrics were similar between 49% and 70%.

An indicator for the feasibility in different inquiry contexts is the answer to the question whether the agreement between the ranking of the students and the ranking of the researchers is more or less equal in each three successive inquiry units. In summary, there is a 91% agreement in rankings for the rubrics that were used in the science unit. The rankings in

[1] Although this rubric was filled out by only three of the twelve student groups during the biology unit, the rankings matched for 100% with those of the researchers.

the rubrics of the biology unit were similar for 64% and those for the physics unit for 86%. A more detailed analysis shows that the largest differences were seen in the similarity in the rankings of the rubrics about *hypothesis* (nr. 3), *research method* (nr. 4), *evaluation of accuracy* (nr. 16) and *evaluation of reliability* (nr. 17). Regarding the biology unit there was less than 50% agreement between the rankings of students and researchers in these four rubrics, but in the two other inquiry units there was 72% or more agreement. Remarkable was that whenever differences in rankings appear, about 80% of this disagreement was caused by students giving themselves higher rankings than the researchers did.

We also analysed whether all nineteen rubrics were meaningful in the three successive inquiry contexts. The observations of the lessons and analysis of the inquiries of the students showed that sixteen rubrics were meaningful in at least two successive inquiry units. The rubrics on *taking of a sample* (nr. 5), *performing observations* (nr. 10) and *defining patterns in results* (nr. 17) were actually appropriate in one inquiry unit.

Analysis of the student responses in the questionnaire shows that about half of the students answered that they made quite regular use of the benchmark samples while applying the rubrics. Some quotes of students: “The examples were useful in understanding the descriptions [in the rubric] and help to check your own work.” “The examples are about different science topics than the inquiry was about. Sometimes it was about biology while I did an inquiry in physics.” “I didn’t understand some of the examples, for instance those that dealt with spread. Too complex. And it was not practical that some examples were only visible after following the link in the digital document.”

Connection with prerequisite knowledge of students

We also analyzed whether the pre- and unistructural level of the self-evaluation instrument were in line with the student prerequisite knowledge about the used CoE. For ten of the twelve student groups the first rankings in the rubrics *experiment: independent variable* (nr. 8) and *comparability of results* (nr. 13) were on the assumed relational level. The other two student groups were ranked in both rubrics on the multistructural level. Also was observed that nine of the twelve student groups performed on multistructural level when they formulated a *research question* (nr. 2) and *draw a conclusion &* [made] *use of evidence* (nr. 14) in the first inquiry unit. For an overview of these findings, see Table 5.

Table 5. Rubrics with too elementary descriptions on pre- and unistructural levels

Rubric (number and title)	
2	Inquiry question
8	Experiment: independent variable
13	Comparability of results
14	Draw a conclusion & use of evidence

Actual hierarchy of levels in the instrument

Although the levels in the designed rubrics were supposed to be described hierarchical as deduced from the SOLO taxonomy, use in class should show whether these levels were experienced as hierarchical. When the researchers filled out the rubrics they both established that in thirteen of the nineteen used rubrics the descriptions on the different levels seem to be more or less hierarchically built. In six rubrics (see Table 6) the unistructural, multistructural and relational levels showed to be interchangeable when filling out the rubrics. Because of this inconsistency in the design of the instrument, the researchers mostly ranked students’ inquiry methods on a particular level without meeting the requirements of the previous level(s).

The students and teachers also observed the lack of hierarchy in some of the rubrics. One student wrote down on her questionnaire: “[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”. During the interviews, one student was more positive on the hierarchy in the rubrics: “It can be seen from the rubrics on which level you formulate and conclude certain things”. On the subsequent question whether he actually recognized different levels in all rubrics he answered: “Mostly I understood what the difference was between the descriptions, and sometimes I filled out what matched my inquiry best, and negotiated some [parts] of the description. It was mostly elucidated by the examples”. Another student disagreed with this opinion on the benchmark samples: “Sometimes I didn’t see why one example was better than the other. I thought: the examples are reversed by accident.”

Table 6. Rubrics with interchangeable levels as indicated by the researchers

Rubric (number and title)	
4	Research method of an experiment or observation
6	Preparation of tables to note down data
8	Experiment: independent variable
9	Experiment: dependent variable
11	Mean & spread of measurement values
13	Comparability of results

Discussion

This study was done to evaluate which design characteristics are needed to design a self-evaluation instrument that is feasible for pre-university science students to evaluate the accuracy, reliability and validity in successive science inquiry units. We formulated four design characteristics that were implemented in a self-evaluation instrument with nineteen rubrics (see Table 1). On basis of the findings of our study, we now can reflect on whether these design characteristics actually led to a feasible instrument for self-evaluation by pre-university science students.

As described in the theoretical perspective, the analytical trait of the instrument *(design characteristic 1)* seems to be useful for pre-university science students, because they have less experience in evaluating the accuracy, reliability and validity of an inquiry and need specific feedback to improve their performance (Arter & McTighe, 2001). Nevertheless, providing the students with a set of analytical rubrics to evaluate the accuracy, reliability or validity was not enough to learn which CoE contribute to these aspects of an inquiry. As shown in Table 4, students did not make use of all rubrics they were provided with and stated that they could not evaluate the parts they had not done. It seems that the students only evaluated the CoE that they applied during the preparation, performance and completion of their inquiries. In a next inquiry unit, they still did not make use of those CoE, although they had seen in their set of rubrics that they could use it. In our view, this means that a self-evaluation instrument for novices in evaluating the accuracy, reliability and accuracy should partly have an analytical and somewhat a holistic character. Extending the instrument with a holistic part fits in with learning of novices. It looks like they should first get the scheme of evaluating accuracy, reliability and validity of an inquiry in mind, before they can transfer these ideas to other learning contexts (Bransford, 2000).

A second reflection can be made on the generic character of the instrument and the contribution of the benchmark samples to this aspect *(design characteristic 2)*. In our analysis on the agreement between the rankings of the researchers and the student groups, we saw that most students’ rankings were more or less similar to those of the researchers, but there was a difference in the percentage of similarity in the different inquiry units. It appeared that during the biology inquiry unit the students had faced more difficulties in the evaluating the accuracy, reliability and validity of their inquiry correct than in the other two inquiry units, as was visible in the lower agreement between the students' rankings and the rankings of the researchers. From our own experience as teachers we know that students sometimes have more information in mind than they write down and one could say that these differences in ranking could be caused by the amount of information the students wrote down.

However, there is a different cause for the dissimilarities in rating in this study, because the lack of written information from the student groups should have appeared in the other inquiry units and led to less similarity in ranking, too. A more plausible explanation can be deduced from the use of the benchmark samples by the students. A lot of students seem to make use of the examples to better understand the generic descriptions in the rubric. Probably for the students the benchmark samples are easier to apply on an inquiry in a physics context than in a biological context and lead to more agreement in the respective rankings in a physic inquiry context.

We also observed that three rubrics are in our study only meaningful in one of the inquiry units. These were the rubrics on *taking a sample* (nr. 5), *performing observations* (nr. 10) and *defining of patterns in results* (nr. 15). Although these CoE can be meaningful for students in other inquiry units than those in our study, it has to be reconsidered whether they should be part of the redesigned self-evaluation instrument. Students can be confused when provided with a rubric about a CoE that is not applicable in their inquiry. Novices have not yet the flexibility to see whether a CoE fits in the 'pattern' of their inquiry (Bransford, 2000). As a consequence, the students can apply a CoE on their inquiry as is visible in the following student response "because I have a rubric about this CoE".

Regarding design characteristic 3 the students did make use only of 14 out of the 19 rubrics they were provided with (Table 4). In our view, a reason for this might be the absence of some CoE in the inquiries of the students. Moreover, the students were less experienced in doing inquiries and are used to perform just the major parts of an inquiry they explicitly asked for in the learning materials (e.g. formulating inquiry question, drawing a conclusion). The five CoE that are scarcely or never used, seem to be not needed for novices who are evaluating the accuracy, reliability and validity of an inquiry.

Nevertheless, each of these CoE is important to be part of the self-evaluation instrument, but it can be questioned whether all these CoE have to be incorporated in rubrics. The instrument can consist of more components than rubrics and information about the more advanced CoE can be provided in the learning materials of the inquiry units.

Finally, we reflect on the actual hierarchy of the levels in the self-evaluation instrument and on the question whether the descriptions of these levels lead to a feasible instrument for pre-university science students *(design characteristic 4)*. It was observed and found in interviews that students seem to be motivated by the self-evaluation instrument to achieve higher levels of performance.

As shown in Table 5, the descriptions at pre- and unistructural levels in most rubrics seem to match the prerequisite knowledge of the students. Four rubrics appear to have too low-skilled unistructural levels for pre-university science students, in spite of the evaluation in a previous study of the use of CoE by pre-university science students (Van der Jagt, Schalk, & Van Rens, 2011).

The majorities of the rubrics seem to have multistructural, relational and extended abstract levels that are built in a hierarchical way on the unistructural way, in accordance with the SOLO taxonomy. Six rubrics appear to have a lack of hierarchy (Table 6) which could be caused by the differences between the application of strategies in an inquiry and actions executed to fulfill an inquiry (Lunetta, Hofstein, & Clough, 2007). Strategies can be applied on different levels, e.g. a student can formulate an inquiry question with no variables, one appropriate (in)dependent variable, a lot of inappropriate variables and so on. Actions on the other hand cannot be described on different levels, because someone fulfills it or not. For example, a student repeats his measurements or he does not. It is rarely seen that students only repeat one of the measurements. As a matter of fact, it looks like CoE that corresponds with strategies can be described on (five) different levels of performance while actions can only be described as present or absent.

Conclusions

From the findings and discussion it can be concluded that the self-evaluation instrument with rubrics has the potential to be a feasible instrument for students in evaluating the accuracy, reliability and validity of an inquiry. The analytical trait of the instrument *(design characteristic 1)* is a necessary condition for pre-university science students to help them to identify the CoE that contribute to the accuracy, reliability and validity of an inquiry. Nevertheless, the students also have a need for a more holistic instrument to get an overview about the connection between CoE and accuracy, reliability and validity in an inquiry.

About the expected generic character of the instrument *(design characteristic 2)* it can be concluded that the majority of the rubrics has the potential for evaluating the accuracy, reliability and validity in different inquiry contexts, but the content of at least four rubrics is not yet properly formulated to reach this aim. The benchmark samples are supportive for the students, but are not always interpreted as generic as expected by the students.

The student groups regularly made use of fourteen rubrics of the self-evaluation instrument, mostly rubrics that focus on the validity of an inquiry *(design characteristic 3)*. The other five CoE are scarcely used by pre-university science students in doing inquiries and the accompanying five rubrics seem to be too advanced in a self-evaluation instrument for novices in doing inquiries.

A final conclusion can be drawn on the hierarchical characteristic of the rubrics in the self-evaluation instrument *(design characteristic 4)*. The pre- and unistructural level of fifteen rubrics seem to have an evident connection to the prerequisite knowledge of pre-university science students. Four rubrics should be improved to meet this design characteristic (Table 6). Thirteen rubrics seems to have descriptions on multistructural, relational en extended abstract levels that are built to a more or less extent in a hierarchical way on the unistructural level. The other six rubrics have interchangeable descriptions.

Implications: Revision of the self-evaluation instrument

Based on the previous discussion and the conclusion an improved self-evaluation instrument in inquiries will contain ten rubrics (see Table 7). These ten showed to be major analytical steps for pre-university science students in evaluating the accuracy, reliability and validity of an inquiry. Other motives for maintaining these rubrics are 1) the high similarities between the rankings of the students and the rankings of the researchers, which means a high reliability as a self-evaluation instrument; 2) a satisfactory feasibility in different inquiry contexts; 3) the possibility of use by the students in different inquiry units; and 4) the items from the CoE-model are or can actually be described in a hierarchical way on the five levels of the SOLO taxonomy. If necessary (see Table 5 and 6), minor revisions on the formulation and hierarchy of the descriptions should be made, especially when descriptions are a mixture of actions and strategies. Furthermore, the ten rubrics need a major revision on the content of the benchmark samples so as to enlarge the usability in different inquiry contexts. The benchmark samples should all belong to the same inquiry context and be more univocally to the students.

Table 7. Rubrics to be maintained in the revised self-evaluation instrument

Rubric (number and title)	
1	Theoretical perspective
2	Inquiry question
3	Hypothesis
5	Taking a sample
11	Mean and spread of measurement values
14	Drawing conclusion & use of evidence
16	Evaluation of accuracy of measurements
17	Evaluation of reliability of the results
18	Evaluation of validity of conclusion
19	Recommendations for supplementary inquiries

In our opinion, the other nine items from the CoE-model are still valuable to be learned to students as part of learning how to evaluate the accuracy, reliability and validity of inquiries. These CoE should appear in a more holistic way in the revised self-evaluation instrument or in other learning materials of the different inquiry tasks. Rubrics that in this study are only filled out by students in one inquiry unit, for example the rubric about *performing observations* (nr. 12), can be integrated in the work sheets of this particular inquiry unit and be explained by the teacher. In this way students learn why this CoE is important for the validity in a particular inquiry, but will not be confused by its meaningless appearance as part of the self-evaluation instrument in other inquiries. Another option is to provide students with an easy-to-use checklist to control their performance before they apply a set of rubrics for evaluation purposes.

Based on the outcomes of this study we revised the design characteristics (see Table 8). In summary, a self-evaluation instrument with rubrics seems to have the potential that pre-university science students learn how to evaluate the accuracy, reliability and validity of an inquiry, when it also contain a holistic tool by which students can get an overview of CoE that are important for the accuracy, reliability and validity of an inquiry.

Table 8. Overview of revised design characteristics of a student self-evaluation instrument for evaluating the accuracy, reliability and validity of an inquiry.

Design characteristics for the revision of the self-evaluation instrument	
1.	The instrument is composed of rubrics that have an analytic trait and is accompanied by a tool that gives a holistic overview of the connection between CoE and the accuracy, reliability and validity of an inquiry.
2.	The instrument is generic for evaluating the accuracy, reliability and validity in inquiry units in the different school science subjects. It contains benchmark samples to elucidate the generic descriptions. The benchmark samples are formulated around the same subject and serve as examples for the whole range of inquiry contexts.
3.	The instrument contains a set of 10 rubrics that are based on 10 CoE which are connected with strategies.
4.	Each rubric has five hierarchical levels conform the SOLO taxonomy. Pre- and unistructural level are based on the prerequisite knowledge of pre-university students. Multistructural, relational and extended abstract levels are built in a hierarchical way on the unistructural level.
5.	The nine CoE which are related with actions are, whenever possible, integrated in the tool with a holistic trait (see design characteristic 1).

Appendix I

The content of the rubric section *Inquiry question* (nr.2)

Rating 🡻*Mark the description that best fits your inquiry question*	**Level of description**	**Example**
1	The inquiry question is formulated rather general. The formulation is mostly based on knowledge from your daily life	*What is liquor?*
2	In the inquiry question, you mention one of the variables of the inquiry or you mention more than one independent and/or dependent variable. You make use of professional terms to formulate the research question.	*What happens to your blood rate when you're standing upside down?*
3	In the inquiry question, you mention the independent and dependent variable of the inquiry. The formulation shows that you have basic knowledge of the research issue.	*Which washing-up liquid cleans the best: one with zeolite of one with phosphates?*
4	In the inquiry question you give, with the help of the relevant variables, an explicit description of the objective of the inquiry. The formulation 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	In the inquiry question it is shown that you want to use this experiment to enlarge scientific knowledge in the research field. The formulation 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?*

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 Practica, 48*(1), 12-19.

Arter, J., & McTighe, J. (2001). *Scoring Rubrics in the Classroom*. Thousand Oaks, California: Corwin Press, Inc.

Biggs, J., & Tang, C. (Eds.). (2007). *Teaching for Quality Learning at University* (3rd ed.). Buckingham: Open University Press.

Bransford, J. D. (2000). *How People Learn: Brain, Mind, Experience, and School* (Expanded ed.). Washington, D.C.: National Academy Press.

Chinn, C. A., & Malhotra, B. A. (2002). Epistemologically Authentic Inquiry in Schools: A Theoretical Framework for Evaluating Inquiry Tasks. *Science Education, 86*, 175-218.

Cohen, L., & Manion, L. (1994). *Research Methods in Education* (4th ed.). London: Routledge.

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/Philadelphia: Open University Press.

Gott, R., & Duggan, S. (2003). *Understanding and Using Scientific Evidence*. London: SAGE Publications.

Gott, R., Duggan, S., Roberts, R., & Hussain, A. (n.d.). Research into Understanding Scientific Evidence. Retrieved May 19, 2009, from http://www.dur.ac.uk/rosalyn.roberts/Evidence/cofev.htm

Hodges, L. C., & Harvey, L. C. (2003). Evaluation of Student Learning in Organic Chemistry Using the SOLO Taxonomy *J. Chem. Educ. , 80*, 785.

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, California: SAGE.

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

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

Lake, D. (1999). Helping students to go SOLO: teaching critical numeracy in the biological sciences. *Journal of Biological Education, 33*(4), 191-198.

Ledford, B. R., & Sleeman, P. J. (2000). *Instructional Design: A Primer*. Greenwich, Conneticut: 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: 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, New Jersey: Lawrence Erlbaum Associates.

Mertler, C. A. (2001). Designing Scoring Rubrics for Your Classroom [Electronic Version]. *Practical Assessment, Research & Evaluation*, 7. Retrieved March 30, 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: 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.

Moskal, B. M. (2000). Scoring Rubrics: What, When and How? [Electronic Version]. *Practical Assessment, Research & Evaluation*, 7. Retrieved April 2, 2009 from http://PAREonline.net/getvn.asp?v=7&n=3.

Moskal, B. M., & Leydens, J. A. (2000). Scoring Rubric Development: Validity and Reliability [Electronic Version]. *Practical Assessment, Research & Evaluation*, 7. Retrieved 30 March 2009 from http://PAREonline.net/getvn.asp?v=7&n=10.

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.

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: Beta 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. Retrieved March 30, 2009 from http://PAREonline.net/getvn.asp?v=9&n=2.

Van der Jagt, S., Schalk, H., & Van Rens, L. (2011). *Teachers' and Students' Use of Concepts of Evidence in Judging the Quality of an Inquiry.* Paper presented at the Authenticity in Biology Education: Benefits and Challenges. A selection of papers presented at the 8th Conference of European Researchers in Didactics of Biology (ERIDOB), Braga, Portugal.

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, Published online Wiley InterScience*.

Yin, R. K. (2003). *Case Study Research: Design and Methods* (3rd ed.). Thousand Oaks, California: SAGE.