Development of a student self-evaluation instrument to evaluate accuracy, reliability and validity (a/r/v) in inquiries

Saskia van der Jagt[1,3], Lisette van Rens[1], Herman Schalk[1], Albert Pilot[2] and Jos Beishuizen[1]

[1]Faculty of Psychology and Education, VU University Amsterdam, The Netherlands; [2]FISME, Utrecht University, The Netherlands; [3]Coornhert Gymnasium, Gouda, The Netherlands

Abstract

This educational design study aims at operationalising design characteristics that lead to a pre-university science students' feasible self-evaluation instrument to evaluate the accuracy, reliability and validity (a/r/v) in successive science inquiry units. A self-evaluation instrument with nineteen rubrics was designed, based on four characteristics that were identified from the literature, which included 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. It is concluded that the self-evaluation instrument with rubrics seems to have the potential in teaching pre-university science students how to evaluate the a/r/v of an inquiry. A major revision regards that part of the students self-evaluation instrument which needs to become holistic instead of analytic.

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

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 (a/r/v) 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). 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 a/r/v 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 this aspect (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 a/r/v (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 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). A

possibility for teaching pre-university science students how to evaluate the a/r/v 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 assess the content of student products rather than processes or strategies. More particularly, it is not known which characteristics of rubrics can help to improve the strategies of students in ensuring the a/r/v 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 a/r/v in successive science inquiry units?

Theoretical perspective: Design characteristics of a self-evaluation instrument with rubrics

In the 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. These 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) and will provide the design of a feasible instrument that students can use to evaluate the a/r/v in an inquiry.

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. Holistic rubrics are seen as a means to make an overall judgment about the quality of a task. Analytic rubrics are also useful in giving specific feedback to students and for self-evaluating purposes (Arter & McTighe, 2001; Mertler, 2001). Students with less experience in performing a specific task, in our case evaluating the a/r/v of an inquiry, learn the most from using rubrics with an analytical trait. Therefore, for the aim of our study, we opt for a set of analytic rubrics *(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 (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. 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 of the authors in chemistry and biology education 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 a/r/v of an inquiry (Gott & Duggan, 2003). This suggests that the content of a self-evaluation instrument regarding a/r/v of an inquiry can be related to the items in the CoE-model that are connected to the a/r/v of the inquiry design, the actual measurements, 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 a/r/v 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

Intended use Subject	**Intend to evaluate**
After preparing the inquiry	
Theoretical framework	Validity
Inquiry question	Validity
Hypothesis	Validity
Research method of an experiment or observation	Reliability
Taking of a sample	Reliability
Preparation of tables to note down data	Validity
Preparation of handling & analysis of data	Validity
After collecting the data	
Experiment: independent variable	Validity
Experiment: dependent variable	Reliability
Performing observations	Accuracy
Mean & spread of measurement values	Accuracy
After handling the data	
Handling of outliers in measurement values	Accuracy
Comparability of results	Reliability
Drawing conclusion & use of evidence	Validity
Defining of patterns in results	Validity
After evaluation of the inquiry	
Evaluation of accuracy of the measurements	Validity
Evaluation of reliability of the results	Validity
Evaluation of validity of the conclusion	Validity
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. To show the successive steps in evaluating the a/r/v of an inquiry, all descriptions and benchmark samples in the rubrics should be represented hierarchical and should be easy to be distinguished for students (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 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).

The SOLO taxonomy uses five levels: prestructural, unistructural, multistructural, relational and extended abstract. The prestructural and unistructural levels of the SOLO taxonomy 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 a/r/v 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. 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)*.

Methodology

To evaluate the feasibility of the designed instrument, a qualitative research method was used with a triangulation of data (Yin, 2003). The instrument is tested 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 a/r/v of an inquiry.

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

Data collection and analysis

In each of the three inquiry units the student groups were asked to evaluate the a/r/v of their inquiries with rubrics. 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: 1) the inquiry plans, data sets, conclusions and discussions of the student groups; 2) the rankings of the student in the rubrics from three inquiry units; 3) videotapes of the teachers' instructions on the use of the rubrics; 4) questionnaires about the use of the instrument immediately after the students had completed each of the three inquiry units; 5) the opinions on the use of the instrument of four students, who were interviewed after completion of all three inquiry units; 6) 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 a/r/v of student group inquiry plans, data sets and conclusions and discussions were rated independently by two researchers - with the same self-evaluation instrument as was used by the

students (inter-rater reliability: 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 a/r/v 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 the multistructural, relational and extended abstract level were hierarchical built on the unistructural level. All data were independently analysed by two researchers and discussed until consensus was reached (Janesick, 2000).

Findings

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*, *inquiry question*, *hypothesis* and *research method of an experiment or observation* in the successive inquiry units. The students scarcely or never filled out the other rubrics. 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"*. 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* and on *drawing the conclusion & use of evidence*. Now and then, a student pair evaluated the *handling of outliers*. The rubric *Defining patterns in results* is never used by the students 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*, *evaluation of reliability* and *evaluation of validity*. Half of the student pairs filled out the rubric on *recommendations for supplementary inquiries*. Analyses of the observations in classroom and the video recordings support these comments.

Analysis of the student responses in the questionnaire regarding the actual use of the rubrics reveals responses like: *"Half of the rubrics I could not use, because I had not done these things during my inquiry"*. One teacher stated that students first had to learn which steps they have to make during the performance of an inquiry: *"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."*

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"* and *"The examples are about different science topics than the inquiry was about. Sometimes it was about biology while I did an inquiry in physics."*

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 a/r/v 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 inquiries since high agreement would indicate that students were able to use the instrument as intended.

Analysis of the feasibility of the instrument showed 80% or more agreement in the ranking of the students and researchers in the *theoretical framework*, *inquiry question*, *taking a sample* , *drawing conclusion & use of evidence*, *evaluation of reliability* and *evaluation of validity*. Less than 40% agreement between the rankings was seen in the rubrics *research method of an experiment of observation* and *evaluation of accuracy*. 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 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 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*, *research method*, *evaluation of accuracy* and *evaluation of reliability*. 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. Whenever differences in rankings appear, about 80% of this disagreement was caused by students giving themselves higher rankings than the researchers did.

Actual hierarchy of levels in the instrument

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 the unistructural, multistructural and relational levels showed to be interchangeable when filling out the rubrics. These were the rubrics on *research method of an experiment or observation, preparation of tables to note down data,* e*xperiment: independent variable, experiment: dependent variable, mean & spread of measurement values* and *comparability of results.* 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 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".*

Discussion

This study was done to evaluate which design characteristics are needed to design an instrument that is feasible for pre-university science students to self-evaluate the a/r/v in successive science inquiry units. On basis of the findings of our study, we now can reflect on whether the four identified 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 a/r/v 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. The 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 a/r/v should partly have an analytical and somewhat a holistic character.

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 a/r/v 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. Many 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*, *performing observations* and *defining of patterns in results*. 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. 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 a/r/v 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.

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 a/r/v 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 Concepts of Evidence (CoE) that contribute to the a/r/v 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 a/r/v 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 a/r/v 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 last conclusion can be drawn on the hierarchical characteristic of the rubrics in the self-evaluation instrument *(design characteristic 4)*. 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 2). These ten showed to be major analytical steps for pre-university science students in evaluating the a/r/v of an inquiry.

Table 2. Rubrics to be maintained in the revised self-evaluation instrument

Rubric
Theoretical perspective
Inquiry question
Hypothesis
Taking a sample
Mean and spread of measurement values
Drawing conclusion & use of evidence
Evaluation of accuracy of measurements
Evaluation of reliability of the results
Evaluation of validity of conclusion
Recommendations for supplementary inquiries

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, 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.

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 a/r/v of inquiries. These CoE should appear in a more holistic way in the revised self-evaluation instrument.

Based on the outcomes of this study we revised the design characteristics (see Table 3). In summary, a self-evaluation instrument with rubrics seems to have the potential that pre-university science students learn how to evaluate the a/r/v of an inquiry, when it also contain a holistic tool by which students can get an overview of CoE that are important for the a/r/v of an inquiry.

Table 3. Revised design characteristics

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 a/r/v of an inquiry.
2. The instrument is generic for evaluating the a/r/v 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 ten rubrics that are based on ten CoE which are connected with strategies. The nine CoE which are related with actions are, whenever possible, integrated in the tool with a holistic trait (see design characteristic 1).
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.

References

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

Arter, J., & McTighe, J. (2001). *Scoring Rubrics in the Classroom*. Thousand Oaks, 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.

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.

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

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.

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.

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.

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. In A. Yarden & G. S. Carvalho (Eds.), *Authenticity in Biology Education: Benefits and Challenges. A selection of papers presented at the 8th Conference of European Researchers in Didactics of Biology (ERIDOB)* (pp. 41-52). Braga, Portugal: CIEC, Universidade do Minho.

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