THE USE OF CONCEPTS OF EVIDENCE BY STUDENTS IN BIOLOGY INVESTIGATIONS: DEVELOPMENT RESEARCH IN PRE-UNIVERSITY EDUCATION

Herman H. Schalk, Joop A. van der Schee,
Onderwijscentrum VU, CETAR Centre for Educational Training, Assessment and Research,
Vrije Universiteit Amsterdam,
and
Kerst Th. Boersma,
FIsme, Freudenthal Institute for Science and Mathematics Education, Utrecht University
h.schalk@ond.vu.nl

Paper presented at the 7th ERIDOB Conference,
September 16-20, 2008,
Utrecht University, The Netherlands

Abstract

Learning to ensure the quality of investigations requires the understanding of empirical evidence. For this purpose *concepts of evidence* (CoE) were introduced to the students in two versions of a teaching and learning strategy. 50 students from the 11th grade (age 16-17) of two urban high schools carried out open investigations in biology. In addition to oral feedback, 25 students received written feedback during their investigations twice. The other 25 students carried out tasks in which they reflected upon the CoE in other investigative contexts.

Focus in the evaluation of the strategy was the extent to which the student reports showed the use of CoE. Also, a written test for the use of CoE was administered before and after the investigation projects.

The group as a whole performed better than the students of six other high schools, which were mostly 12th-graders. Also, their results on the post-test were significantly higher than those on the pre-test. Within the group of 50, students who did the reflection tasks performed better on the test, but worse in their reports compared to students who were not invited to reflect upon the CoE outside their own investigations. Therefore, it seems fruitful to combine the strengths of both versions: explicit, written feedback on the use of CoE in their own investigations together with recontextualizing the concepts in the context of other investigations.

1. Introduction

1.1 Learning about science

Student investigations in biology education may serve different goals. They may aim at expanding knowledge about biological subjects, like the factors effecting the composition of the macro fauna of a pond by making an inventory of the species and relating them to conditions of light and nutrients. This aim is what Abd-El-Khalik *et al.* (2004) call 'inquiry as means' or 'inquiry *in* science'. But the purpose of an investigation could also be to learn about doing investigations and doing them well. In the same example emphasis is then put on taking samples at relevant locations and of relevant size or on multivariate analysis. That is called 'inquiry as ends' or 'inquiry *about* science'.

1.2 Understanding of evidence

For understanding the way scientific claims are underpinned with evidence, one needs to have 'procedural understanding', which can be described as "the understanding of a set of ideas (…) related to the 'knowing how' of science and concerned with the understanding needed to put science into practice. It is the thinking behind the doing." (Gott & Duggan, 1995, p. 26).

Procedural understanding has been worked out in a set of 'concepts of evidence' (CoE), which have been described in detail by Gott *et al*., (2008) and cover all aspects of scientific research, from the level of a single measurement up to the level of societal issues (fig. 1). In their description of what they mean by 'concepts of evidence', Gott and Duggan (1995) emphasize that the concepts do not refer to the skill of, for instance, taking measurements, but to the decisions that have to be made about what measurements to take, how and how many.

In a recent article, Gott and Duggan (2007) argue that the circles in the diagram of figure 1 can be 'read' from inside out as well as from outside in (fig. 2). In doing an investigation, an investigator builds up his argument from data to claim; in evaluating a claim, a scientific literate citizen tries to see through from claim to data. Therefore, mastering the CoE is helpful – if not a prerequisite – for 'successfully' carrying out investigations as well as for evaluating scientific claims from others.

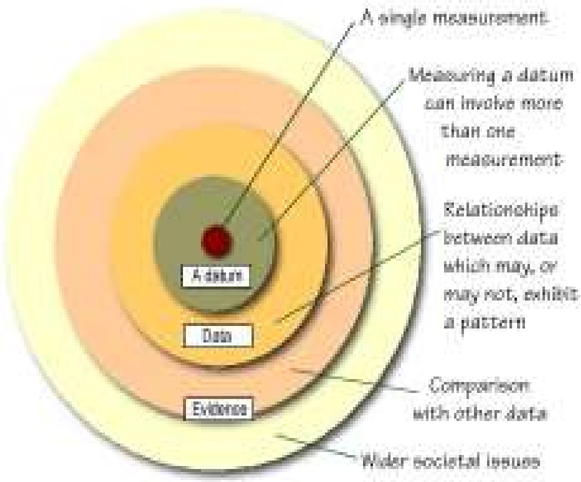

Fig. 1 A framework for data and evidence (from Gott *et al*., 2008).

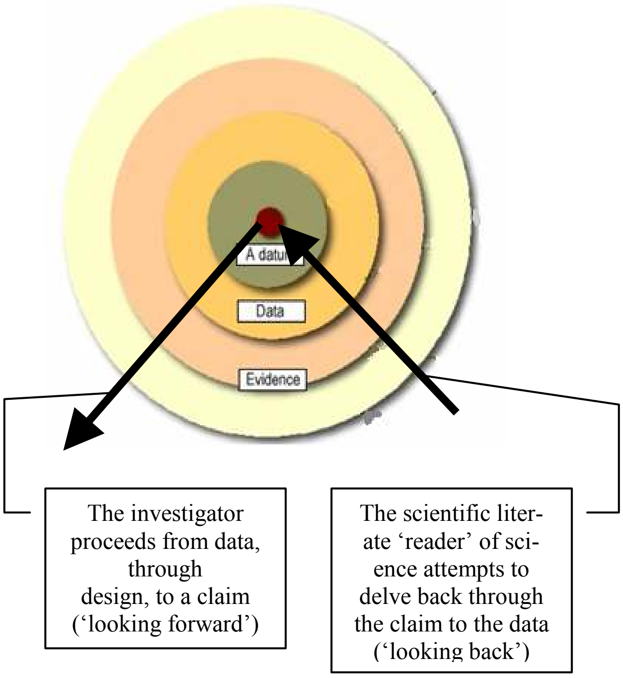

Fig. 2 Looking forward and looking back (after Gott & Duggan, 2007)

1.3 Elaboration of the concepts of evidence

Gott *et al.* (2008) describe the CoE in a very extensive way. They admit that this is not directly applicable in secondary education and they also supply a subset 'appropriate to GCSE science in the UK'. Even then, to apply the CoE to the situation of Dutch biology education, some elaboration is necessary. If the concepts of evidence are to be useful in Dutch biology education, they should fit all or at least most of the types of student investigations in biology.

Table 1 Twenty-three concepts of evidence

Related to the research question

1. A research question:
   - consists only of unambiguous terms and formulations;
   - is sufficiently specific and confined.
2. It is important to distinguish whether the research question demands a description or the testing of a hypothesis.

Related to the hypothesis

3. A hypothesis should be formulated that fits the research question: it is an expected result or a possible explanation.
4. The hypothesis should be testable.
5. It is (nearly) always possible to postulate more than one hypothesis.
6. Based on the hypothesis (and a number of assumptions), a prediction is formulated about which observations or measurements can be expected (if the hypothesis [and the assumptions] is true, then …).

Related to the design of the enquiry

7. In descriptive studies, the features to be observed should be stated explicitly.
8. In hypothesis testing research, the dependent and the independent variables should be identified.
9. In hypothesis testing research, all other variables that may have an influence should be identified and kept constant if possible.
10. In hypothesis testing research, a 'control experiment' should be included.

Related to observations and measurements

11. Observations and measurements should not influence the outcome of those observations or measurements themselves, and any unavoidable influence must be made explicit.
12. The range and the intervals of the chosen values of the independent variable should match the expected variation of the dependent variable.
13. The sensitivity of the instrument should be appropriate to the measurements to be taken.
14. The sample should be representative of the population
    - large enough
    - random or stratified, and the method of sampling should have nothing to do with the subject of research.

    The number of measurements or observations should be large enough.
15. Aberrant or anomalous results should be examined to decide whether they should be used further (can they be qualified as outliers?).
16. If measurements are averaged, it should be in accordance with the content.
17. One should only speak of differences between measurements if the difference is statistically significant.

Related to conclusions and explanation

18. The conclusions should match the research question and (in the case of hypothesis testing) the hypothesis.
19. A correlation is not the same as a causal relationship.
20. It should be stated whether the explanation is causal (related to the causes) or functional (related to the consequences).

Related to evaluation

21. If an enquiry is not carried out validly and reliably (in other words: if the preceding conditions are not met), the results are open to dispute.
22. If possible, the results of the enquiry should be confirmed by data from other research.
23. The conclusion should be compared with accepted ideas (theories), common sense and experience.

One type of investigation that is common in Dutch biology education is the descriptive study of for instance an ecosystem or the behaviour of animals in a zoo. When given the opportunity to formulate questions of their own students are not only interested in relations between variables, but also in the features of the natural world. In that respect 16-17 year olds do not differ very much from younger children (Keys, 1998). Application of the concepts should therefore not be restricted to the evidence from hypothesis testing investigations, but also cover the design and evaluation of descriptive studies. Therefore, we have attempted to reformulate the concepts in a such a way that descriptive as well as hypothesis testing enquiries are included. Another adaptation we have made, is reformulating them as a set of requirements, criteria, for any inquiry. A draft of our version of the CoE was presented to several experts in theoretical biology, biology teachers and science education researchers. Comments resulted in a set of 23 concepts (see table 1), not all of which will be applicable to every investigation, but in our opinion every investigation can be evaluated with a subset.

1.4 Learning by doing

In his contribution to the above cited article (Abd-El-Khalik *et al.*, 2004), Lederman stresses the difference between doing inquiry and knowing about inquiry. He states: “My biggest fear for the future, however, is that science teaching will continue to focus on the performance of inquiry skills to the exclusion of understandings about inquiry and NOS *[Nature of Science]*, .. (…) In particular, it is often believed that students will develop understanding about scientific inquiry and NOS simply by doing science.” (p. 403). He cites ‘extensive research’ that indicates that “students do not come to understand either inquiry or NOS as a consequence of having experienced scientific inquiry or inquiry-oriented classroom climates.” Also Millar *et al.* (1994) and Gott and Duggan (1995) assert that the understanding of evidence requires a ‘body of knowledge’ which has to be taught explicitly.

This view is in contrast to those expressed by e.g. Hodson and Hodson (1998) and Lave (1997). By way of – respectively – ‘enculturation’ and ‘cognitive apprenticeship’ “apprentices learn to think, argue, act, and interact in increasingly knowledgeable ways with people who do something well, by doing it with them as legitimate, peripheral participants” (Lave, 1997; p. 19). Scientific inquiry is imbedded in a social context with it’s own culture and speech (Mortimer & Scott, 2003) and students should participate in ‘authentic’ inquiry in order to become skilled researchers. Science and scientific inquiry aren’t primarily a body of knowledge but an activity, a practice. These last two terms refer to cultural historic activity theory. In the light of this theory (see for a description e.g. Van Eijck (2006), p. 16), the concepts of evidence serve as tools to conduct investigations in the same way as they are conducted in scientific practice and the concepts should be acquired in an authentic context. This leads to the conclusion that *learning about science* cannot take place without *doing science.*.

However, the above cited observation of Lederman still stands: doing science does not automatically lead to understanding science. It requires a teaching and learning strategy that is explicitly designed for that purpose. That such a strategy must include ‘hands-on’ activities does not mean that ‘brains-on’ activities can be skipped.

1.5 Authenticity

Students in school often like to do investigations, but their main purpose is mostly not to find an answer to their research question, but to get a good mark. Students working on an investigation are often very proud of their results. They are very much committed to the investigation, they conceive it as something of their own, they really want to make something good out of it. After working hard on writing down everything they have found out in their investigation and making a nice presentation they expect their teacher to be very content as well. But how disappointed they are when they read her comment: “Your report did not convince me

that cat grass grows best under green light. The evidence you have gathered is too weak to draw that conclusion." This situation can be found in many Dutch high schools. Students work hard, but the quality of their work does not always meet the standards of 'good research'. It lacks understanding of criteria for gathering and evaluating the quality of empirical evidence, the already mentioned 'procedural understanding' (Millar *et al.,* 1994). This problem not only exists in the Netherlands (Smits *et al*., 2000; Van der Schee & Rijborz, 2003), in other countries as well it has been observed (Duveen *et al.,* 1993; Chin & Brown, 2000; Keys & Bryan, 2001; Tiberghien *et al.,* 2001).
And that is not the only difference between the practice of science and the practice of school science. Teachers aren't scientific researchers, and as a rule the purpose of the investigations is not the discovery of new knowledge but the confirmation of what is already known. At best, scientific practice in school context is a 'didactisized' practice with – in part – its own rules, habits and *speech genre* (Mortimer & Scott, 2003).

1.6 Recontextualisation

Yet another important difference between school practice and authentic practices of scientific inquiry is the temporarily limitedness of the first. Only limited time in school can be spent on doing investigations, often it is limited to only one more or less open investigations in a complete school career. And one can pose the (rhetoric?) question whether one investigation is enough to develop understanding of evidence needed to ensure the quality of subsequent investigations and to be able to evaluate scientific claims made by others.
Indeed, to develop an ability to use concepts in different contexts 'recontextualization' seems necessary (Van Oers, 2001). That is, by using concepts in different contexts – in their context-specific meaning – a certain degree of abstraction can be reached, facilitating the use of them in new contexts.
The answer to this problem seems easy: let students do more than one investigation and recontextualize the CoE from one investigation to another. But in school there isn't always time to do that and – as stated before – not every concept is relevant in every investigation.
In the team of teachers with whom we designed our teaching and learning strategy, we discussed about the question how to stimulate the development of (a certain selection of) the CoE in one open investigation project. In this case, 'open' means that students choose their subject and research question themselves, so that it cannot be predicted which of the CoE will be relevant in the context of their investigation. In other words, the discussion was if we could 'guarantee' that all students would encounter all of the CoE we considered important.
In this discussion, we did not come to consensus. Some teachers were of the opinion that students are only open and committed to learn about these concepts if they are of direct relevance to their own investigations; addressing other concepts would be a waste of energy. Others were inclined to add extra tasks to ensure that all the CoE we selected to be addressed were indeed addressed, independent of the subject of their own investigations.
We decided to try both, so we developed two versions of the strategy. In the 'implicit' version the development of understanding of evidence was stimulated by giving oral and written feedback on plans and draft reports of the investigations with a focus on the use of our selection of CoE. In the 'explicit' version, besides oral feedback, the students were prompted to interrupt their own investigations to carry out four reflection tasks in which the selection of the CoE was addressed in the context of other investigations.

1.7 Research questions

To summarize: with respect to the understanding of evidence two related goals were identified: (a) students should learn to ensure the quality of their own investigations, and (b) stu-

dents should learn to evaluate the quality of investigations of others. The purpose of our research is to develop an effective and feasible strategy for reaching these goals.
The research questions are:

- Does the developed teaching and learning strategy help students in pre-university biology education to better ensure the quality of their own investigations? *(Is the strategy good enough?)*
- Can students use the CoE in evaluating investigations of others?
- Is there a difference between the two versions of the strategy in respect to the above questions?

The answer to the first question comes from observations and the analysis of their products. The second question is answered by evaluating the use of CoE in students' answers on a written test. The third question asks for a further analysis of the answers on the first two.

2. Research design and method

2.1 Research design

2.1.1 Development research

Since the purpose of this study was (and is) to develop a feasible and effective teaching and learning strategy (TLS), it can be characterized as development research or design research. The main characteristics of this approach are that the research is interventionist, iterative and oriented towards process, utility and theory (Van den Akker *et al.*, 2006). In our study this means that we are not following a (quasi) experimental design with an experimental and a control treatment, but that we are evaluating two 'treatments' with regard to their contribution to the development of understanding of evidence. Figure 3 gives an outline of the design of the study.

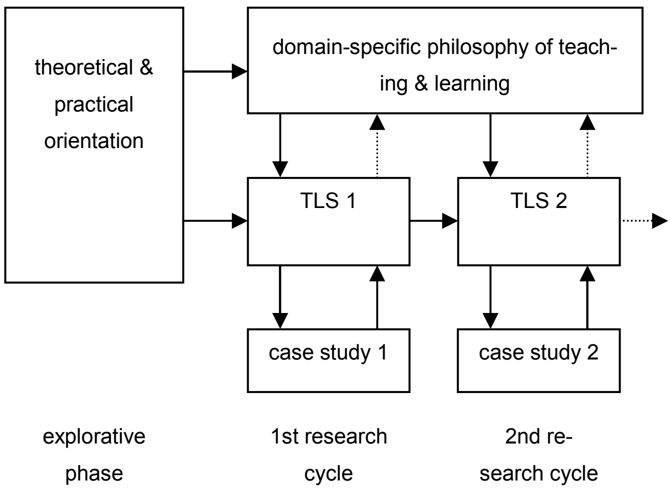

Fig. 3. Design of the study

2.1.2 Explorative phase

Besides a theoretical orientation on learning about science and on doing science, a practical orientation was carried out with regard to the use of CoE in curriculum documents, textbooks, teaching and student investigations (Schalk, 2006). This included the analysis of 169 student investigation reports of six high schools in the Netherlands with regard to the use of CoE .

2.1.3 Design of the teaching and learning strategy

In cooperation with six experienced biology teachers, we developed a first draft of the TLS. After a first cycle of testing and evaluating we formulated a set of design principles for the second version:

- Students can work on their investigations in class during nine lessons (40-45 minutes); besides they can carry out parts of their investigations at other moments, in school, at home or outside. Only in class they can ask questions to their teacher.

- These lessons are preceded by an introductory lesson to provide a common basis about what should be understood by ‘good inquiry’, based upon what the students already know.
- Students are provided with a booklet to support their investigations; the booklet contains a description of various types of investigations and our selection of the CoE (concepts number 1, 3, 4, 5, 6, 8, 9, 10, 14, 18 and 21, see table 1).
- Students work in small groups of two or three; the groups of students choose their own topic, formulate their own research question and make their own research design. The subject of the investigation can be anything within the scope of biology.
- During the lessons in class the teachers stimulate reflection on the CoE by discussing the students’ investigations, either by initiating discussion themselves or in reaction to students’ questions.
- In the *implicit version of the TLS* the students receive written feedback at the level of their own investigations twice. In these comments the teachers incorporate remarks related to the CoE, especially to the selected concepts.
- In the *explicit version of the TLS* students are asked to pause their own investigations four times to carry out a reflection task. Each task presents a problem, asks students to give a solution and to explain why it is a solution, and contains a question about which general criteria for good research (concepts of evidence) can be derived from the given problem and its solution. The first task is about how to formulate a good research question (concepts nr. 1 and 18), the second about hypothesis testing (concepts nr. 3, 4, 5 and 6), the third about controlling variables and taking samples (concepts nr. 8, 9, 10 and 14). In the fourth task students are invited to give comments on draft versions of each other’s research reports, by using CoE (especially nr. 21).
- In both versions, students report the design and results of their investigations in a ‘scientific’ report, for which they are given guidelines in their booklet. Draft versions are commented upon by the teacher (in the implicit version) or by fellow students (in the explicit version).

In our design of the TLS we have tried to approach a scientific practice as much as possible in the specific situation of the schools and teachers involved.

2.1.4 Students

Our study was carried out in two urban schools for pre-university education, in each school with an experienced biology teacher. The students were in the 11th grade, age 16 or 17. In the second research cycle, there were a total of 50 students, 32 girls and 18 boys. Besides the subject of biology, nearly all students in our study also did mathematics, physics and chemistry. In each school half of the students followed the explicit version and the other half the implicit version.

2.1.5 Test

A written test for the use of CoE was developed. It consisted of 31 items in which students were given descriptions of several aspects of investigations done by others. They were asked to analyse and evaluate those investigations or to give suggestions for improvement. The CoE involved in the test were the same as those on which emphasis was put in the TLS.
The test was administered before and after the students carried out their investigations. There was two to three months time between pre- and post-administration of the test.

2.2 Analysis

2.2.1 Student reports

The 169 student investigation reports from the explorative phase of our study and the 23 student reports from our second research cycle were analyzed in the same way. For every of the 23 CoE – interpreted as criteria to be met – each product was analyzed and a score was given in on of the following four categories:

(1) the criterion is not applicable (this is possible for some of the CoE)

(2) the criterion is met

(3) the criterion is not met, but there is an argumentation why the criterion is not met (e.g. due to practical circumstances).

(4) the criterion is not met, and there is no argumentation why the criterion is not met.

If a score in category (2) or (3) was given, the conclusion was that the report showed understanding of evidence with regard to this concept. We determined the proportion of the number of cases in which understanding was shown to the number of cases in which it could have been shown, that is:

$$\{\Sigma(2) + \Sigma(3)\} / \{\Sigma(2) + \Sigma(3) + \Sigma(4)\} \times 100$$

The outcome of this formula is the extent to which a certain concept is 'understood' by the students, in terms of a percentage. The analysis of the final reports of the 23 student investigations (done in small groups by 50 students in total) was carried out by two independent researchers (proportion agreement: 0.81; Cohen's kappa: 0.71), discussion until consensus.
The results for the separate CoE were grouped and averaged in three categories: *Research question, hypothesis and prediction* (concepts 1-6), *Design, observations and measurements* (concepts 7-17) and *Conclusion and evaluation* (concepts 18-23). Also, results were calculated for the concepts in our selection of CoE.

2.2.2 Test

Answers to the test were scored with a code indicating if a concept of evidence was used, and if so, which. If the concept used was relevant for the question one or more points were given for the answer. This provides information about the degree of understanding of evidence the students show as well as about the specific concepts they use to evaluate others' investigations.
All tests were analysed by two independent researchers (proportion agreement 0.82), after which results were discussed until consensus. Scores were analysed using the software TiaPlus® (build nr. 300), a program for analysing tests and items. On the basis of this analysis seven items were removed, which resulted in a test of 24 items, with a maximum score of 27.
A number of students scored considerably lower on the post-test than on the pre-test. The teachers' impression was they did not work seriously on the post-test. These students' scores were removed from the analysis, after which the scores of 41 students remained in the analysis.
Test scores were determined for the test as a whole as well as for three subtests reflecting the above mentioned three categories of CoE. Psychometric analysis showed a Crohnbach's alpha of 0.59. This is low, in our view mainly due to its restricted length – 24 items – and its internal diversity; the estimation of alpha if the test consisted of 40 items is 0.70, for the subtest this estimation is even higher, which of course indicates a greater internal coherence. After all, Crohnbach's alpha is an indication of a test's internal coherence. Verhelst (2000) emphasizes that alpha (just as other ways of determining the reliability of a test) is only an estima-

tion of the lower bound of the reliability. He presents different ways of determining the reliability of a test after only one administration. One of these is the estimation of the Greatest Lower Bound (GLB), which with small groups gives an estimation that is too high. He shows that with large groups (min. 200) the GLB approaches asymptotically the reliability. For our test as a whole the GLB is 0.91, which clearly is an overestimation. In table 2 the reliability of the test and its subtests is shown.

Table 2 Reliability of the test and three subtests calculated in different ways.

	number of students	*number of items*	*max. score*	*Crohnbach's alpha*	*alpha if test consisted of 40 items*	*GLB*
test	41	24	27	0,59	0,70	0,91
subtest 1	41	10	12	0,53	0,82	0,84
subtest 2	41	7	8	0,46	0,83	0,67
subtest 3	41	7	7	0,33	0,74	0,54

3. Results

3.1 Reports

As shown in Table 3, the investigation reports of our students showed more understanding of evidence than the reports we analyzed in our exploration phase. There is not a huge difference, but the fact should be taken into account that the latter reports were mainly from 12th graders at the end of their schooling, whereas our students were in the 11th grade.
The differences in the subset 'conclusion and evaluation' are the largest: 52 to 43 percent for the concepts 18-23 and 83 to 72 when it is restricted to the elements 18 and 21 (on which was put extra emphasis). The 'implicit' group performs better than the explicit (with 56 and 86 percent in these categories). Therefore, we prudently conclude that our strategy is 'good enough' in learning to ensure the quality of investigations (our first research question) and that the 'implicit' version seems slightly better.

3.2 Test

Table 4 shows the results of the pre- and post-test. It can be seen that most post-test scores are significantly higher than the pre-test scores (paired-samples *t* test, $p<0.05$). The differences are greater in the 'explicit' group. Because the focus of the test is the use of CoE in the evaluation of other investigations, we also conclude that our strategy enhances this ability (our second research question) and that the 'explicit' version seems slightly better in this respect.

Table 3 Proportions of understanding of evidence (realised compared to possible) in the final reports in the different groups and the orientation. *Top*: all concepts; *bottom*: the selected concepts.

	explicit (N=12)	*implicit (N=11)*	*total (N=23)*	*orientation (N=169)*
Research question, hypothesis and prediction (1-6)	66	76	71	67
Design, observations and measurements (7-17)	67	76	71	64
Conclusion and evaluation (18-23)	49	56	52	43
All concepts	63	72	68	61
Research question, hypothesis and prediction (1, 3, 4, 5, 6)	74	80	77	74
Design, observations and measurements (8, 9, 10, 14)	57	72	64	64
Conclusion and evaluation (18, 21)	79	86	83	72
All selected concepts	69	78	74	70

Table 4 Average p'-values (fraction of maximum score) on the test and the three subtests in pre- and post-test. *: significant difference between pre- and post-test (*p* < 0,05*)*.

	explicit (N = 20)		*implicit (N = 21)*		*total (N = 41)*	
	pre-test	*post-test*	*pre-test*	*post-test*	*pre-test*	*post-test*
Whole test (max. 27)	0.51	0.66*	0.53	0.61*	0.52	0.63*
subtest 1 (max. 12)	0.66	0.78*	0.72	0.80*	0.69	0.79*
subtest 2 (max. 8)	0.36	0.58*	0.35	0.44	0.35	0.51*
subtest 3 (max. 7)	0.44	0.53	0.40	0.47	0.42	0.50*

4. Conclusions and implications

So, the first and second research question are answered positively: our TLS seems good enough to learn students to use CoE in their own investigations as well as in evaluating others' investigations. With respect to the third research question we conclude that the 'implicit' version of our strategy yields better results with regard to the quality of the investigations of the students themselves and that the 'explicit' version leads to a better use of CoE in evaluating other's investigations.

If we try to explain these results in terms of our design principles, we may – prudently, of course – conclude that *doing* investigations in combination with instruction and feedback in which emphasis is put on concepts of evidence leads to students products that show - in our opinion - a sufficient amount of understanding of evidence. Of course this is dependent upon the way in which feedback was given and reflection upon the CoE was stimulated. The teachers' role in putting the focus on the CoE on the descriptive, the explanatory as well as the generalized level appeared to be crucial, as we reported elsewhere (Schalk *et al.*, 2007), but this became possible because students were engaged in and committed to investigations of their own, that is, in *doing science*.

The authenticity of the student investigations can be discussed, of course. The investigations took place in school practice and not in an authentic science investigation practice like a research institute or a university. And the guidance came from teachers that had some research experience (which seems essential; Windschl, 2003), but were no researchers themselves. Most students were highly committed to their investigations and were eager to find the results, although it cannot be denied that there were also students for whom the mark they would get was more important. Therefore, if it is not authenticity, it might be commitment that is crucial for learning to ensure the quality of investigations.

That the students who had most time in class to spend on their own investigations and received written feedback from their teachers (the 'implicit' group) were better in applying the CoE to their own investigations isn't surprising. Also, that the reflection tasks in the 'explicit' group, in which CoE were applied to other investigations, led to better results on the test isn't surprising either. Therefore, it seems fruitful to combine the strengths of both: oral as well as written feedback on the use of CoE in their own investigations. Feedback of the students on the implicit version learned they appreciated this very much. But this should be combined with the reflection tasks to learn to recontextualize the CoE in other contexts (Van Oers, 2001).

The implications of these conclusions for the development of understanding of evidence seem to be the following.

- Stimulate commitment and authenticity by letting students carry out open investigations in which they can choose subject, research question and design themselves.
- Put emphasis on relevant concepts of evidence in introduction and in oral and written feedback.
- Present the concepts of evidence in more than one context. Let students reflect upon the quality of their own investigations by taking a step aside: not only afterwards, but also during the investigation process itself.

REFERENCES

Abd-El-Khalick, F., BouJaoude, S., Duschl, R., Lederman, N.G., Mamlok-Naaman, R., Hofstein, A, Niaz, M, Treagust, D., Tuan, H.-L. (2004). Inquiry in science education: International perspectives. *Science Education* 88 , 397-419.

Chin, C., & Brown, D.E. (2000). Learning in science: a comparison of deep and surface approaches. *Journal of Research in Science Teaching* 37, 109-138.

Duveen, J., Scott, L. & Solomon, J (1993). Pupils' understanding of science: description of experiments or 'A passion to explain'? *School Science Review* 75 (271), 19-27.

Gott, R., & Duggan, S. (1995). *Investigative Work in the Science Curriculum.* Buckingham / Philadelphia: Open University Press.
Gott, R. & Duggan, S. (2007). A framework for practical work in science and scientific literacy through argumentation. *Research in Science and Technological Education* 25, 271-291.
Gott, R., Duggan, S., Roberts, R., & Hussain, A. (2008). *Research into understanding scientific evidence* (1-14). www.dur.ac.uk/rosalyn.roberts/Evidence/cofev.htm. Durham: School of Education.
Hodson, D., & Hodson, J. (1998). Science education as enculturation: some implications for practice. *School Science Review* 80 (290), 17-24.
Keys, C.W. (1998). A study of grade six students generating questions and plans for open-ended investigations. *Research in Science Education* 28, 301-316.
Keys, C.W., & Bryan, L.A. (2001). Co-constructing inquiry-based science with teachers: essential research for lasting reform. *Journal of Research in Science Teaching* 38 (6), 631-645.
Lave, J. (1997). The Culture of Acquisition and the Practice of Understanding. In D. Kirshner, & J. Whitson, J. (Eds.), *Situated Cognition: Social, Semiotic, and Psychological Perspectives* (pp.17-35). Mahwah NJ: Lawrence Erlbaum Associates.
Millar, R., Lubben, F., Gott, R., & Duggan, S. (1994). Investigating in the school science laboratory: conceptual and procedural knowledge and their influence. *Research Papers in Education* 9, 207-249.
Mortimer, E.F., & Scott, Ph.H. (2003). *Meaning Making in Secondary Science Classrooms*. Maidenhead: Open University Press.
Schalk, H.H. (2006) *Zeker weten? Leren de kwaliteit van biologie-onderzoek te bewaken in 5 vwo* [*Are you certain? Learning to ensure the quality of biology research in pre-university education*]. Amsterdam: Onderwijscentrum VU.
Schalk, H.H., Van der Schee, J.A. & Boersma, K.Th. (2007). *The development of understanding of evidence in pre-university biology education in the Netherlands*. Paper presented at the 7th ESERA conference, August 21-25, Malmö, Sweden.
Smits, Th., Lijnse, P.L., & Bergen, Th. (2000). Leerlingonderzoek met kwaliteit. [Student research with quality] *Tijdschrift voor Didactiek der β-wetenschappen* 17 (1), 14-30.
TiaPlus®: this software package can be obtained at CITO for non-commercial, scientific use (www.cito.nl).
Tiberghien, A., Veillard, L., Le Marechal, J.-F., Buty, C., & Millar, R. (2001). An analysis of labwork tasks used in science teaching at upper secondary school and university levels in several European countries. *Science Education* 85, 483-508.
Van den Akker, J., Gravemeijer, K., McKenney, S. & Nieveen, N. (2006). Introducing educational design research. In: Van den Akker, J., Gravemeijer, K., McKenney, S. & Nieveen, N. (eds.) (pp.3-7). *Educational design research*. London and New York: Routledge.
Van der Schee, J.A., & Rijborz, J.D. (2003). Coaching students in research skills: a difficult task for teachers. *European Journal of Teacher Education* 26, 229-237.
Van Eijck, M.W. (2006). *Teaching quantitative concepts with ICT in pre-university biology education; the case of datalogging the heart*. PhD thesis, Amsterdam University.
Van Oers, B. (2001). Contextualisation for abstraction. *Cognitive Science Quarterly*, 1, 279-306.
Verhelst, N.D. (2000). *Estimating the reliability of a test from a single test administration*. Measurement and Research Department Report 98-1. Arnhem: Cito.
Windschitl, M. (2003). Folk Theories of "Inquiry": How Preservice Teachers Reproduce the Discourse and Practices of an Atheoretical Scientific Method. *Journal of Research in Science Education* 41 (5), 481-512.