www.sails-project.eu

Report on the assessment frameworks and instruments for IBSE skills - Part B

SAILS
Strategies for Assessment of Inquiry Learning in Science

Strategies for Assessment of Inquiry Learning in Science

D2.3 Report on the assessment frameworks and instruments for IBSE skills - Part B

Authors: Csaba Csíkos, Benő Csapó, Gábor Veres, Ferencné Adorján, Katalin Radnóti
Project name: Strategies for the Assessment of Inquiry Learning in Science (SAILS)
Project number: 289085
Start date: 01/01/2012
Duration: 48 months
Lead partner for this deliverable: University of Szeged
Project coordinator: Dublin City University
Contact: info@sails-project.eu
Website: www.sails-project.eu

The research leading to these results has received funding from the European Union's Seventh Framework Programme for research technological development and demonstration under grant agreement no 289085
This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both. This document does not represent the opinion of the European Union, and the European Union is not responsible for any use that might be made of its content.

Contents

Introduction

A key aim of the SAILS project is to present a framework for the assessment of inquiry learning in science. The purpose of this framework is to provide a detailed description of the content of assessment and to describe what and how to assess in the context of IBSE. In this work the frameworks may be used as models, in particular as a way of presenting assessment content and examples for the assessment items.

In the earlier report D2.2, the structure of the proposed framework for SAILS was outlined. As developing the complete framework will take time, the structure was operationalised in order to develop the initial assessment items for piloting and trialling. Three topics were selected that provided teachers with a view of the type of assessment opportunities that could be developed within each of the subject areas (physics, chemistry and biology) and provided specific examples of what and how to assess inquiry skills. The topics chosen for the initial assessment items were Food, Rates of Reaction and Speed aimed at the lower second level students and these topics were chosen so as to be applicable across most countries. These initial assessment items were presented in D2.2 and a report on their trialling with teachers has been reported in D3.1.

From this analysis, it was clear that we had to develop a common understanding of particular inquiry skills within the consortium and also to show in more detail the strategies used in assessing particular skills. Therefore, several draft units were developed by members of the consortium, trialled by different partners, and reported on in the form of case studies. Following further evaluation and testing with teachers, these draft units and case studies will be combined into SAILS UNITS that give details of good inquiry lessons with embedded assessment modes and criteria. These draft units and case studies have been analysed to inform this framework document.

This document is in two sections – the first part focusses on the understanding of particular skills in inquiry, while the second section presents the criteria used in the assessment of particular skills.

1. Objectives to be assessed in IBSE contexts

Assessment of IBSE skills and competencies requires teachers to be able to use a variety of tools to determine where students are in their learning. From these data, they can make judgements that can help the student to decide on the next step in learning, and so guide them towards improvement (this data can be used in both formative ways and also in summative ways – therefore the distinction in terms of formative and summative modes of assessment is not important in this context.

While there are a range of skills and competencies developed through inquiry, within the context of the current document, the focus is on how content knowledge, inquiry skills, reasoning ability and scientific literacy may be assessed within an inquiry lesson. Each of these aspects is discussed below.

1.1 Curricular content

Inquiry-based science education must fulfil the criteria of providing appropriate scientific knowledge. As the participating countries of the SAILS project differ in school structure, curricula requirements and assessment practices (see Deliverable 1.2), we focus here on some common and important scientific concepts, and how the assessment of conceptual understanding was addressed in different case studies in lessons. In general, a reversed order of instructional phases is suggested in line with what mathematics education call the transition from DTP (definition-theorem-proof) to PTD (prooftheorem-definition) phases in mathematics lessons. Analogically, scientific concepts (their definition and connections) can be addressed at the beginning of a lesson, then a scientific phenomenon or law can be introduced, and finally an empirical demonstration might occur. This "traditional" sequence should sometimes be replaced by the PTD-analogue where empirical investigations comes first, then having summarized the data and presented the results, the necessary conceptual development should take place.

There has been a world-wide shift from traditional teacher-centred pedagogical approaches to student-centred active learning. This results in changes in the teaching of scientific concepts, for example through inquiry processes. One particular approach considers two phases of the science classes of particular importance in developing students' conceptual understanding. Firstly, when approaching a problem or planning an experiment at the beginning of the lesson, the necessary prerequisite knowledge components may be evoked through discussions/brainstorming activities. Students are usually allowed and encouraged to freely use any terms they think to be related to the current problem. This brainstorming or mind mapping collection of ideas often reveals students' misconceptions. The most commonly used scientific terms such as speed, heat, impetus, or life have a rich semantic framework developed from students' past experiences and former studies. Secondly, having conducted active observations and experiments, students reflect on their findings using appropriate scientific terms. In many cases, due to the experiences gathered, student themselves feel the need to use more precise terms at this stage of the learning process.

In the literature on scientific misconceptions, the debate on how to replace previous naïve beliefs and misconceptions with scientific conceptions has in recent decades been turned to an even more powerful idea: in many cases, students can consciously restructure their conceptual network, or – an equally feasible option – they can separate the semantic structures of the same word. Momentum, impetus or dynamism can be used interchangeably in everyday conversations, but through investigations within a physics lesson students learn to connect the words they use in everyday life to other, scientifically defined words and to their mental images. Another example of restructuring or separating concepts is "life" which in everyday contexts is closely associated with moving or breathing, and through experiences provided by active learning contexts in the school, finally, the word "homeostasis" may be appropriately built in students' semantic networks.

Therefore when we assess students' content knowledge, we pay special attention to the introductory and to the concluding phases of a lesson. The case studies provided by consortium members will provide examples of how a teacher can facilitate learning and assess the use of scientific terms.

1.2 Reasoning skills and abilities

In Deliverable D2.1 "Report on the Strategy for the assessment of skills and competencies suitable for IBSE", reasoning skills necessary for scientific inquiry were identified and described, e.g., deductive (logical) reasoning, inductive reasoning, combinatorial and probabilistic reasoning. There are widely known diagnostic assessment devices for these reasoning processes. In classroom settings, however, the "rational errors" of thinking may be observed and corrected. For instance, the misuse of some plausible implication schemes (denial of antecedent or affirmation of consequent rules) may lead to incoherent conclusions. The structure of two premises (p implies q, and q is true; or p implies q and "not p" is true) tends to tempt students to erroneously conclude that p or "not q" is true, respectively.

When separating variables of an experiment, keeping one or more variables constant while changing others, an appropriate level of combinatorial reasoning is crucial. For example, if two variables are to be manipulated, it is essential to understand that there are at least four possible situations. Although combinatorics is an abstract mathematical phenomenon, in inquiry tasks its use is content- and context-bound, and instead of using an abstract combinatorial skill, many students fail to list the possibilities arising from the number of variables they define.

1.3 Scientific literacy

PISA studies have highlighted the importance of thinking processes that go beyond the mere recall of factual knowledge and the immediate use of routine algorithms. Scientific literacy incorporates the idea of individual and societal usefulness of knowledge, therefore the knowledge components learnt in school should be transferrable to new contexts and for several different purposes. One commonly cited example of a "science literacy task" is the Semmelweis' diary task released immediately after the first PISA survey in 2000. Typically scientific literacy tasks are connected to relatively long, real-life texts, requiring students to distinguish important and distracting data, actively use their prior factual knowledge and reasoning skills developed through both formal and informal learning. Scientific literacy is particularly important within an inquiry class when finding problems and posing relevant questions, and when considering the generalizability of the findings in an experiment. Of course, in other inquiry phases, e.g. in planning an investigation, knowledge transfer occurs as a result of finding analogies between real-life experiences and the constraints of the current problems, the literacy components of scientific knowledge are utilised.

1.4 Inquiry skills

In the SAILS project, the following definition given by Linn and Davis (2004) is followed. This definition has been further discussed and elaborated in a WP1 milestone draft of the SAILS project (Draft report on key skills and competencies).

> *Inquiry is the intentional process of diagnosing problems, critiquing experiments, and distinguishing alternatives, planning investigations, researching conjectures, searching for information, constructing models, debating with peers, and forming coherent arguments.*

The *genus proximum* of this definition is "intentional process", while the *differentia specificae* are the inquiry skills. Firstly, the use of the term *intentional* refers to conscious, strategic processes that may and should be generalizable throughout different domains and contexts. Secondly, the term *skill* can be reliably used since the acquisition of these processes are different from knowing merely when or how to use them. It was decided to focus on developing assessment strategies connected to common curriculum topics. Even where inquiry skills are dealt with directly (e.g., a lesson on how to combine independent and dependent variables, how to define control samples) it is assumed that students have prior experiences from experiments conducted in different domains and contexts.

It would be desirable to form an agreement in the scientific discourse community about a coherent and complete list of inquiry skills. Other authors, like Fradd, Lee, Sutman, and Saxton (2001), and Wenning (2007) suggested other well-defined classifications of inquiry skills. These two sources have been used in previous reports of the SAILS project by WP1 and WP2. There is one common feature behind all these attempts; namely, they follow the sequence of inquiry phases. A widely recognized model of inquiry phases is called the five Es model (see e.g., Bybee, 2009) in which five keywords structure the inquiry processes: engagement, exploration, explanation, elaboration, evaluation.

From the Fradd et al. (2001) and Wenning papers (2007), a comparative table can be derived. This table simplifies both taxonomies in order to make them comparable. The main message of this table is: taxonomies of inquiry skills are and necessarily should be anchored to the consecutive phases of scientific inquiry.

Table 1 A comparison of two widely recognized taxonomies of inquiry skills.

"Wenning-skills"	**"Fradd-skills"**
Identify a problem to be investigated.	questioning
Formulate a hypothesis.	
Design experimental procedures to test the prediction.	planning
Conduct a scientific experiment; collect meaningful data, organize, and analyse data accurately and precisely.	implementing
Apply numerical and statistical methods to numerical data to reach and support conclusions.	concluding
Using available technology, report, display, and defend the results of an investigation to audiences that might include professionals and technical experts.	reporting
	applying

From these lists, four inquiry skills were chosen with the purpose of illustrating assessment strategies for inquiry for a number of reasons. Firstly, these skills are often addressed and assessed within the case studies developed through the trialling of draft units by experienced inquiry teachers in each of the SAILS partner countries. Secondly, these four skills can be considered as representative of the different clusters of inquiry skills provided in the above-mentioned papers. A third perspective of focusing on some skills comes from the frameworks of international educational assessment surveys.

The importance in assessing and fostering inquiry skills in classroom situations is highlighted by the fact that both the PISA and TIMSS surveys specifically represent those skills in their frameworks, and consequently, in their tasks. A comparative analysis of the TIMSS and PISA frameworks (Mullis et al., 2009; OECD, 2013) shows the following pattern: both in the fields of mathematics and science, the TIMSS and PISA assessment frameworks address a variety of inquiry skills with the exception of the 'on-site' design and implementation skills. Another important argument in favour of using the PISA and TIMSS frameworks and justifications of focusing on the assessment of inquiry skills may be that the 'formulating situations mathematically' processes in the PISA mathematical literacy study can be compared to the inquiry skills taxonomies of science literacy. It means that initiatives of assessment on inquiry skills in science educations can inform the theory and practice of 21st century mathematics education as well.

The four inquiry skills addressed in this report are: planning investigations, developing hypotheses, debating with peers, and forming coherent arguments. All these skills can be assessed in a way that (1) teachers can diagnose whether students possess an appropriate level of that skill, and (2) teachers can provide feedback and guidance to their students in order to improve students' performance.

1.4.1 Planning Investigations

This skill (called "designing experimental procedures" by Wenning, 2007) refers to the intentional thinking processes necessary before beginning an experiment. Fradd et al. (2001) provide more details about this inquiry skill cluster: (a) decide what you want to do to find out the answer to the question, (b) decide what materials you need, (c) decide how to record the information, (d) decide how to analyse the information, (e) decide how to report the findings. The (a) component of this skill points to the general question how open the inquiry process should be. Fradd et al. (2001) showed that questioning is seldom left as the students' responsibility; therefore students typically react to the question posed by their teacher.

1.4.2 Developing Hypotheses

Theoretically, developing hypotheses may be an inquiry phase that precedes planning investigations. In Fradd et al. (2001), questioning as an inquiry skill comes first and consists of two things: posing questions, and making hypotheses. Wenning (2007) drew our attention to the complementarity of thinking processes that underlie this skill: inductive reasoning is used when formulating hypotheses, and deductive reasoning enables for making predictions from the hypotheses. The reason why we discuss the Planning Investigations skill first is that in the classroom setting the problem or question to investigate often comes from the teacher or from the booklet, and the necessary materials and equipment are also provided. Another reason is that in the case of open inquiry, several new hypotheses may be developed throughout the students' work.

1.4.3 Debating with peers

Observing and assessing the quality of collaborative work is still a great challenge in educational research. The PISA 2012 framework (OECD, 2013, p. 120.) emphasizes that the PISA problem solving survey assessed individual competencies, since there are "significant measurement challenges associated with including collaborative tasks in a large-scale international survey such as PISA". At the classroom level, however, it is still possible and desirable to assess the quality of both individual and collaborative efforts. Both in the Fradd et al. (2001) and Wenning (2007) taxonomies, several phases and skills can be associated with group work. The implementation of an experiment, collecting and analyzing data, and reporting the results all provide opportunities for group work and offer the possibility of assessing the quality of debating with peers.

1.4.4 Forming coherent arguments

This inquiry skill is partly covered by the domain-general reasoning processes described in section 1.2 of this report. There are two reasons for explicitly defining this skill. Firstly, as an inquiry skill, forming coherent arguments is a content- and context-bound intentional process. The quality of argumentation depends on the characteristics of the task (e.g. the presence or absence of prerequisite knowledge, and the cognitive load demand). Secondly, similar to the previous skill, forming coherent arguments (or the lack of it) can be observed and assessed throughout the implementation, analysis and reporting stages of inquiry. This inquiry skill is not explicitly involved in the taxonomy sources; however, reasoning and argument as a phase of the mathematical modelling process is emphasized in the PISA 2012 mathematics framework, and is contrasted with mere mathematical knowledge and skills. The TIMSS 2011 science framework very explicitly claims that students must be able to "construct arguments to support the reasonableness of solutions to problems, conclusions from investigations, or scientific explanations" (Mullis, Martin, Ruddock, O'Sullivan and Preuschoff, 2009, p. 87).

2. Instruments and methods used in the assessment of students' knowledge

In the second part of this report, the knowledge domains and skills are further elaborated and illustrated by examples distilled from real classroom case studies. Table 2 presents the list of the assessed skills in the draft units and case studies and which informed this report. The next WP2 report (Deliverable 2.4) will further elaborate these and will provide a synthesis of all preceding reports. Within the text following, draft units and particular case studies are distinguished as *DU* and *CS#* respectively.

Title	**Developer**	**Case studies**	**Case study title**	**Planning Investigations**	**Developing hypothesis**	**Debating with peers**	**Forming coherent arguments**	**Scientific literacy**	**Scientific reasoning**
Acids, bases, salts	UPRC	UPRC	CS1	DU AND CS	DU AND CS	DU ONLY			
Biotechnology	IEUL	IEUL	CS1			DU and CS	DU and CS		
Black tide: Oil in the water	IEUL	IEUL	CS1	DU AND CS		DU ONLY			
Candle	HUT	HUT	CS1	DU AND CS	DU AND CS	DU ONLY			
Chemical reaction speed	US	US	CS1	DU AND CS	DU AND CS	CS ONLY			
Collision of an egg	US	US	CS1	DU AND CS	DU AND CS		DU AND CS		DU ONLY
Constructing a galvanic cell				DU ONLY	DU ONLY	DU ONLY	DU ONLY		
Cooking an egg		US	CS1	CS ONLY			DU AND CS	DU ONLY	
Cooking food	KCL	KCL	CS1		DU AND CS	DU AND CS		DU AND CS	
Decomposition of starch in saliva	US			DU ONLY					
Electricity	JU	UPJS, HUT, JU1, JU2, DCU	CS1, CS2, CS3, CS4	DU AND CS	CS ONLY	DU AND CS	CS ONLY	DU ONLY	DU ONLY
Fish eating birds	US	US	CS1				DU ONLY	DU ONLY	
Floating orange	KCL	KCL	CS1	DU AND CS		DU AND CS			
Food labels	KCL	DCU	CS1	DU ONLY				DU ONLY	DU ONLY
Galvanic cells	JU	DCU, JU	CS1, CS2	DU AND CS	DU AND CS	DU ONLY	DU AND CS		DU ONLY
Genetic engineering	HKR	HKR	CS1				DU ONLY		
Global warming	HKR	HKR	CS1				DU AND CS		
Goats and Human, resources and sustainability: and the end of the story?	IEUL	IEUL	CS1	DU ONLY					
Height and body mass		US	CS1				DU AND CS		DU ONLY
Household vs natural environment	JU	JU	CS1	DU AND CS	CS ONLY	DU ONLY	DU AND CS		
Martian bacteria in Alentejo	IEUL	IEUL	CS1	DU AND CS		DU AND CS			
Natural selection	SDU	SDU, JU	CS1, CS2	CS ONLY		DU AND CS	DU AND CS		DU ONLY

Plant nutrition	KCL	UPJS	CS1	DU AND CS			DU ONLY		DU ONLY
Polymers	UPJS	UPJS	CS1		DU ONLY				DU ONLY
Rates of change			CS1	DU ONLY	DU ONLY			DU ONLY	
Reaction rates	DCU		CS1	DU ONLY					DU ONLY
Speed	KCL/US	DCU, IEUL, LUH, HUT	CS1, CS2, CS3, CS4	DU AND CS				DU ONLY	
Sports nutrition	HKR	HKR	CS1				CS ONLY		
Temperature of plants	KCL		CS1		DU ONLY				
The probe of the pudding	US		CS1	DU ONLY		DU ONLY			DU ONLY
Ultraviolet radiation	HKR	SDU, LUH	CS1, CS2	DU AND CS	DU ONLY				
Up there… how is it?	IEUL	IEUL	CS1	DU AND CS	DU ONLY		DU ONLY		
Which is the Best Fuel?	HUT	HUT	CS1	DU AND CS	CS ONLY	DU ONLY	DU ONLY		
Wood lice	MaH	DCU, JU, MaH	CS1, CS2, CS3	DU AND CS	DU AND CS		DU AND CS		

2.1 Curricular content

Within the draft units, curriculum content is taught using an inquiry approach. Curricular content is addressed at different levels. In all case studies, a basic scientific concept is taught in a way that includes some inquiry. This is a good example of content-bound development.

The *Speed DU* primarily focused on developing the scientific concept of speed. Speed as the ratio of distance and time might remain a formula in many students' mind, but conceptual development may be promoted through investigations designed by students. Case studies conducted by the consortium partners show that students, through inquiry, were able to build a coherent mental representation of three intertwined concepts: time, speed and distance.

Content elements that are or may be involved in curricula, albeit not of central importance are often evoked by means of brainstorming or mind mapping techniques. In the *Electricity CS3*, the planning skill was addressed by means of mind mapping technique. Here the number of correctly categorized and visualized words used in science and everyday life formed the basis of assessment.

The *Food labels CS1* provided ample opportunities (and this was utilized by means of a brainstorming technique) to recapitulate the conceptual network necessary to talk about food labels. The scientific concepts of energy and the biological (and everyday) semantic network of nutrient, food, junk food were discussed.

In the *Cooking food CS3*, students themselves raised questions worth being investigated. One of these questions simply concerned scientific factual knowledge: "Is cooking a chemical or a physical change?" Such questions gave the opportunity not only to revise some key concepts in science, but to activate the semantic network necessary to formulate investigable research questions on the cooking process.

In the *Ultraviolet radiation CS2*, the factual knowledge on whether water protects from UV radiation served as a preliminary activation of everyday knowledge.

Other content elements may play a marginal role in the curricula; however as "raw material" for experiments they may relate to students' real-life experiences. Nevertheless, these content elements assist the process of development of inquiry skills built around them. For instance, the *Woodlice* case studies focused on biological facts not emphasized elsewhere in the curricula, or the *Cooking an egg CS1* is built around a piece of everyday information.

2.2 Reasoning skills and abilities

The *Chemical reaction speed CS2* explicitly addressed the analogy between the concept of (physical) speed and chemical reaction speed. In order to compare different speeds, students tried to find a solution with the help of analogical thinking. For example, one was the Cooper test (given in a supporting question), another was running 100 m. Through analysing these examples, students realised they could choose between two independent variables - fixed time or fixed distance – while the other would be the dependent variable. Using critical thinking, the students analysed the advantages and the disadvantages of the two alternatives. The answers given by the students could also be used for the evaluation of their performance. Within the case study, this was evaluated during the group discussion but it could also be a part of a students' detailed written assignment.

Correlational reasoning is of huge importance when generating new hypotheses. This reasoning ability refers to the identification of connection between data sets. Independently of whether there is direct causal relation or only a coincidence, students must be able to declare if there is a clear connection between two data sets as described in the *Body mass and height CS1* where data on body mass and height were gathered and hypotheses were formed and examined.

Other reasoning abilities (among which deductive reasoning is almost always implicitly covered by the skill of forming coherent arguments) are partially overlapped by various inquiry skills. While the skill of forming coherent arguments has some developmental potential, formative and diagnostic assessment practices may foster students' thinking by means of revealing their errors and providing feedback on how to improve their thinking.

2.3 Scientific literacy

The assessment of scientific literacy is based on the assessment of its components. As described in section 1.3, several knowledge components are known from and are usable in everyday life. Some good examples from the SAILS case studies illustrate how the assessment of scientific literacy was addressed, e.g. *Fish-eating birds CS1* and *Biotechnology CS1.* Which draft units can be considered as describing everyday authentic problems? The phenomenon of authenticity is culturally-bound, but in Europe the *Food labels*, the *Black tide: Oil in the water*, *Ultraviolet radiation* and *Collision of an egg* draft units are very close to students' interest and activities. When introducing these topics, students feel that they learn something from and about their life.

The assessment of scientific literacy involves questions on scientific inquiry. In other words, inquiry skills are part of scientific literacy when the task is authentic.

2.4 Inquiry skills

2.4.1 Planning investigations

Our examples from the case studies will be grouped into these five clusters described by Fradd et al (2001). Before doing so, note that there can be holistic approaches in assessing planning investigation as an inquiry skill. One example comes from the *Household vs. natural environment CS2* where planning as a whole was assessed on four-level scale. The highest level indicated that the student proposes a consistent and holistic (complete) research plan, and predicts and solves problems that can happen.

Decide what you want to do to find out the answer to the question

The core element of this skill component is the identification, definition and separation of different variables in the experimentation process. One crucial element is whether the student is capable of separating the independent and dependent variables (from 7th or 8th grade, the exact labelling of these variables might be both feasible and desirable). The importance of distinguishing all relevant variables in the experiment can be the basis for assessing the skill of planning investigations. In the *Galvanic cells CS1*, the four level rubric used provides a generalizable idea how this skill may be assessed either during classroom discussion or in post-hoc questionnaires.

0	Nothing
1	Mention concepts from the actual experiment (temperature, plate, etc.)
2	Explicitly state all variables
3	Explicitly state all variables changed and measured

Another highly generalizable idea came from the *Electricity CS3* where an even more extensive scale was used for the assessment of planning inquiry. In the sixth level of that scale, students could list more than 7 things made of different materials for measurement and write down a correct plan of experiment.

Task	Level of execution					
	1	2	3	4	5	6
Planning investigation of conducting properties of different materials	Student can't list things made of different materials for measurement and can't write down a plan of experiment.	Student can list 2-3 things made of different materials for measurement but can't write down a plan of experiment.	Student can list 4-5 things made of different materials for measurement and writes down an incorrect plan of experiment.	Student can list 4-5 things made of different materials for measurement and writes down an almost correct plan of experiment.	Student can list 6-7 things made of different materials for measurement and writes down almost a correct plan of experiment.	Student can list more than 7 things made of different materials for measurement and writes down a correct plan of experiment

Several other examples of assessing this inquiry skill were provided and described in other case studies. In the *Proof of the pudding CS1*, students were required to produce an edible mixture from its two basic components. Students had to decide how to change the ratio of the components. Another variable was the time of cooking. Students' skills of keeping one of these variables constant while manipulating the others have been assessed by means of a three-level rubric.

In the *Cooking food CS3*, a nice example of a four-level rubric is presented for assessing the planning investigation inquiry skill. Of special importance is whether students consider several possible ideas to pursue with high levels of attainment when students "try out several ideas and weighs up which are likely to pursue." The justification may rely on either critical discussion or a scientific explanation.

In the planning phase of the *Chemical reaction speed CS2*, the problem had to be recognised and structured. The starting point was a chemical equation, from which – with the help of some supporting questions - students had to find out what materials to use and how to use these materials in the process.

Some case studies gave students a more open problem where there were many potential variables and it proved to be even more difficult to assess the quality of this skill. In the *Collision of an egg CS1*, students were free to choose dependent variables. One group kept the height of falling constant, and varied the surface while another group decided to try out three different heights with different surfaces. In this latter case not all combinatorial possibilities were consistently tested.

Similarly, in the *Ultraviolet radiation CS1*, students were asked to investigate UV radiation, and they, rather than the teacher, decided whether a lamp in a room or the sunlight would be taken as the source of radiation. The teacher was previously provided with rubrics as assessment tool, but restricted herself to judging students' planning skills based on discussions with students.

The *Woodlice CS1* allowed students to investigate the living conditions of woodlice provided evidence of rather different levels of planning skills. Three different levels were identified, each illustrated in the case study report. One interesting feature of a "mid-level" plan was to add new variables to the experiment inconsistent with the research question instead of eliminating or fixing some of them.

Distinguishing dependent and independent variables is of crucial importance in understanding the planning phase of scientific inquiry. The *Fish-eating bird CS1* provided data about students' difficulties. To the teacher's question "*What were the experimental and the control conditions?"* students often wrongly thought the two phases of the bird's movement in the first experiment were relevant. In some instances the student described the experimental and the control conditions, but did not indicate which was which.

The *Black tide: Oil in the water CS1* and the *Up there how is it CS1* used a three-level rubric to assess the skill of defining the goals of the investigations, with the degree of clarity used in students' descriptions being the defining characteristic used in judging performance. The *Martian bacteria CS1*, however, uses a four-level rubric assessment scale. The lowest level of the four-level scale can be understood as the absence of the inquiry skill, while the highest level refers to a holistic approach in identifying and operationally defining all relevant variables.

The flexibility in planning was highlighted in the *Floating orange CS1*, where students were encouraged to modify their inquiry questions while going through the inquiry.

Decide what materials you need

This component of this skill is rarely addressed in classroom situations. Due to time constraints, teachers usually aim to prepare the necessary equipment and materials. One of those cases where students were free to choose their equipment and materials is the *Collision of an egg CS1* where an extensive range of equipment was available. The teacher even allowed the possibility for students to request additional material.

In the *Woodlice CS2*, students were free to create their environment for the investigation. Collecting all necessary materials and equipment's was a requirement, and differentiation in the assessment began at the point of whether the equipment and materials selected by students allowed for systematic changes of experimental conditions.

The *Black tide: Oil in the water CS1* provided an opportunity to assess knowledge about materials and equipment. The three-level rubrics are based on the selection of all, some or none of the appropriate resources.

The *Martian bacteria CS1* used a four-level rubric for the assessment of this skill component; a low level of performance was assigned for not choosing any resources for the experiment, and the highest level indicated choosing all appropriate resources. The two values in the middle referred to the lack of adequateness or to the incompleteness. This four-level scale seems to be the most highly generalizable when assessing this inquiry sub skill.

Decide how to record the information

In many draft units, teachers' assessment of the appropriate use of equipment was observed. In the *Cooking food CS1*, four-level rubrics (from emerging to extending) were used to assess the level of safe and careful use of equipment. In the *Woodlice CS2*, students' skill in conducting safe and repeatable experiments was scored by dichotomous items. The quality of data collection was scored by a trichotomous item as shown below.

2 points level	4 points level	6 points level
Student can interpret data correctly (categorizing the measured variables as lesser – greater) but cannot create a proper graph based on them	Student can present the data on a graph, but the graph lacks or has poorly developed elements as axes titles, scale, legend etc.	Student can present the data on appropriate graph(s) having all necessary elements as axes titles, scale, legend etc. prepared correctly
Student can point out basic / selected sources of biased / incorrect results of the experiment	Student can enumerate all main factors that might be sources of biased/incorrect results of the experiment	Student can analyse all main factors that might be sources of biased/incorrect results of the experiment and indicate ways to avoid them in the future
Student can propose elements of a method serving to improve the experiment	Student can propose improvement of the course of the entire experiment step by step	Student can compare results of other groups, discuss data interpretation and propose methods to improve both own and the other groups' experiments

In the *Cooking food CS2*, student presentation of data was assessed. In that case, low performance in the four-level rubrics refers to the lack of data tables, while higher levels include the use of tables to present data and the appropriate use of units.

The importance of repeat measurements was emphasized in the *Speed CS1* study. In *Speed CS3*, the teacher felt that a rubric adjusted for self/peer assessment would further improve students' skills with data recording.

Among other factors, the nature of data collection was decided by students in the *Plant nutrition* case studies. The assessment of this skill was done on a three level scale where the decisive characteristics were independence and justification.

In the *Up there how is it CS1*, a three-level rubric was used. Mid-level performance is expressed as adequate performance but lacking some details with a need for reformulation.

Decide how to analyze the information

In the *Chemical reaction speed CS1*, students used Excel diagrams to reveal the connection between reaction speed and temperature. It was an important discovery for the students that not all relationships are linear! After identifying the exponential relationship, relying on the function we could calculate the activation energy needed for the reaction with the help of the some explanation by the teacher. The students were especially pleased to find that such a simple measurement (temperature and time) can lead to so many discoveries based on their theoretical knowledge of physics and chemistry.

In the *Chemical reaction speed CS2*, students had to make decisions on the dependent and independent variables of the experiment. A three-point scale ranked their attainment as weak, average or good. The highest level of this skill presumed not only making the distinction between dependent and independent variables, but giving different possible independent variables.

A question of not merely a technical nature is how and to what extent ICT tools should be used in the data analysis process. Some mathematics curricula include the use of different calculators (including graphing calculators) but teachers' use of ICT often depends on whether such devices can be used within terminal examinations. Some SAILS draft units provide the opportunity to allow students to choose between different data analysis methods including PCs or scientific calculators.

Decide how to report the findings

In the *Chemical reaction speed CS1*, students enjoyed learning how to use Excel to handle their data and were pleased to discover how simple it was to find the function connecting the data, which could be easily read from the trend line. The interpretation and use of R^2 was a novelty to them. Although reporting the results of an experiment is another inquiry skill, in the planning phase making decisions on how to report the findings obviously connects the planning and reporting phases. Therefore the three levels described in the *Household vs. natural environment CS1* for the data presentation skill may serve as a basis for assessing the current skill. At the lowest level, the results are presented only descriptively, and in the mid-level the results are presented in tables and diagrams (appropriate descriptions, axes). In the highest level all other criteria are fulfilled. What is instructive here is that in many scientific experiments, planning an appropriate table format for data collection and then for presentations is a useful idea that should be supported by the teachers.

Planning investigations is multifaceted by nature, and different aspects need to be considered to assess these sub skills. It is logical to define three or four level rubrics for each sub skill, but as is shown in case studies, the assessment rubrics may involve different aspects depending on the task presented to students. For example, being able to describe an experimental plan may be a learning aim of a particular task, while another lesson may involve assessing whether students can determine independent and dependent variables. From the case studies, it seems that it is important for teachers to focus on only two or three particular inquiry skills within a lesson. The aim for the next phase of the project is to provide evidence of assessment practice where immediate or post-hoc feedback is given to progress the development of inquiry skills.

2.4.2 Developing hypothesis

What does a "good hypothesis" look like? Independently of its content, one basic characteristic is that a scientific hypothesis must be – theoretically and practically – falsifiable. Other characteristics can be dependent on the actual content of the task, and one formal feature may sometimes be merely the number of hypotheses developed.

In the *Proof of the pudding CS1*, students had to make some initial predictions about the desired ratio of the pudding components. The quality level of this inquiry skill was assessed by means of three-level rubrics. For this, both the observed classroom discussion and students' answers to a questionnaire were used.

In the *Cooking food CS3*, research questions raised by the students could easily be turned to hypotheses statements. For example, the question "How much water does spaghetti absorb through cooking?" is albeit literally a research question, but it definitely contains identified variables and a statement about the connection of those variables. The observer stated that there was a good deal of diversity in students' research questions, and feedback from teacher about the diversity of research questions is an important assessment tool.

In the *Chemical reaction speed CS1*, an important element of analysis was that the students formulated hypotheses in connection with the predicted temperature dependence of the reaction. Most of the students initially predicted a simple linear relationship. The students were especially pleased to find that such a simple measurement (temperature and time) can lead to so many discoveries based on their theoretical knowledge of physics and chemistry.

In the *Chemical reaction speed CS2*, students were assessed on a three-level scale based on whether they hypothesized a linear relationship between the variables, mention other possibilities, or – using their prior knowledge – they may have assumed logarithmic connection between variables.

Even structured/guided instruction provides opportunities for students to form their own hypotheses. The *Plant nutrition* case studies provided the students with readily defined dependent and independent variables; however, students could vary the actual distance from the source of light and also could revise their initial hypothesis while conducting the experiment.

Students' previous everyday experiences were used in the generating of a hypothesis in the *Collision of an egg CS1*. Students' hypotheses concerned whether the egg remains intact after having impacted on different surfaces.

In the *Ultraviolet radiation CS2*, formulating hypotheses constituted a difficulty for some of the students at the beginning of the lesson. Through observations and a progress report the teacher noticed diverse abilities. Some students were able to formulate hypotheses very systematically and their considerations revealed an appropriate experimental approach. Other student groups needed help from the teacher. They were not able to identify or control variables in an experimental approach.

In the *Woodlice CS1*, three levels of this inquiry skill are illustrated with examples from students' worksheets. An important element of assessing this skill is whether the student developed a falsifiable hypothesis. Those predictions that interchangeably use words with different meanings (e.g., moisture and damp piece of wood) cannot be considered falsifiable therefore indicate a low level of hypothesis making.

Within the case studies, there is debate about the importance of students developing a hypothesis that can be tested. Depending on the aims of the lesson (or depending on the extent to which the teacher's philosophy tends towards content-bound ability development) the feasibility of being able to investigate the hypothesis in the classroom context may be of high importance. In other cases, generating hypotheses not investigable in the classroom but meaningful for scientists, can be appreciated. On one hand, the *Woodlice CS2* used a three-level rubric to assess students' hypothesis development. The main aspect was the feasibility of testing the hypothesis in school.

Assessed skill	2 points level	4 points level	6 points level
Formulating hypotheses	Student can formulate hypotheses that are impossible to be proved by means of a school experiment	Student can formulate hypotheses that after teacher's or colleagues' revision may be proved by means of a school experiment	Student can him/herself formulate hypotheses that may be proved by means of a school experiment

On the other hand, in the *Cooking an Egg CS1*, formulating interesting and relevant questions was more important than the development of testable hypotheses. For example, whether less bacteria die in a shorter time or under lower temperature than more bacteria. (In this case it is impossible to investigate this conjecture under regular classroom conditions; however, it indicates a high level of inquiry skill.)

The *Properties of plastics* case studies highlights student hypotheses on the density, thermal stability and thermal conductivity of plastics. Students completed a questionnaire designed to assess students' metacognition on their understanding of what they did. Surprisingly, one third of the students chose the option "I don't know" to the question "Why did we do it?"

The *Household vs. natural environment CS3* provided a four-level assessment scale for hypothesis development. The highest level of attainment is elaborated as "Student formulates more than two hypotheses; asks research question". An interesting idea may be in a complex problem space to define the quality of hypothesis making in terms of the numbers of hypotheses provided.

Another interesting, and possibly widely generalizable idea is that a three-level scale indicates whether the students was able to produce some elements of hypothesis formulation either by himself/herself or with the help of the teachers, or not at all. This assessment strategy was documented in the *Galvanic cells CS2*.

2.4.3 Debating with peers

Assessment of debating with peers was conducted primarily through teacher observation, after which immediate oral feedback was given to students. Both the quality of the discussion in a group work context and the degree of collaboration was assessed formatively.

In the *Proof of the pudding CS1*, one supportive teacher question was: "How do you consider your own work and that of the groups, what were your strengths and weaknesses?" This question supports not only the quality of group work but also encourages reflective (metacognitive) processes. In the *Cooking food CS3*, criteria for assessing "teamwork" shows some shared features with "debating with peers". The highest level on the four-point rubric contains the term "reporting on progress" thus pointing to some elements of reflective thinking.

In the *Cooking food CS1*, a four-level rubric was used in assessing the quality of group work. It is interesting that there is no "zero" status in the rubrics, since even the lowest quality level labelled "emerging" means that students took part in the group work. The higher levels of this skill refer to more involved/engaged participation, such as taking a role in the group work, discussion and negotiation on group work.

In the *Cooking food* CS2, group work was the focus of assessment. The four-level rubrics were the same as in the Cooking food *CS1*, but this task was tried out with older students. Consequently, the same description of a particular rubric may refer to other cognitive and affective processes. For example, the observer described that students found the open-ended nature of the inquiry to be difficult.

In the *Chemical reaction CS2*, the focus of assessing debating with peers was on the independent, individual questioning that facilitated the work of the group. The distinction between the upper two levels of the three-point scale depended on whether the questions raised were justified or whether appropriate arguments were made.

Evaluation / *Skill*	**Weak**	**Average**	**Good**
Communication, debate / *debating with peers, forming coherent arguments*	Not formulating questions independently, but answering the teacher's questions, rarely forming own opinion	Formulating questions independently, these questions don't always take the work forward, argument not always supported expertly	Formulating inquiry questions independently, being able to build existing disciplinary knowledge and elements of other subjects into arguments

In the *Ultraviolet radiation CS1*, the students ended the lessons with a presentation and a peer discussion on their hypothesis and their methodology. This gave the teacher not only a clear indication of the students' understand of variable control but also the students gained more insight into the processes of planning and carrying out investigations for the future. The skill of planning investigation proved to be an indicator of the overall differences between students.

The *Properties of plastics CS1* provided ample space for collaboration. The teacher assessed the quality of participating in group work by means of self- and peer-assessment, using smiley pictograms. Following the investigation, the teacher reflected that she would adjust the assessment in future classes to provide immediate feedback if problems occurred instead of this post-hoc assessment.

The level of the debating with peers was assessed by means of teacher questions and whole class discussion in the *Food labels CS1*. Besides the quality of argumentation, the individual contribution to the small group discussion formed the basis for assessment.

The *Floating orange CS1* drew attention to the importance of acknowledging not only the verbal communication part of debating among peers but the non-verbal elements (e.g., nodding) must also be considered. The teacher commented that "The difficulty was that some pupils were talking a lot and others only a little but the nods of agreement and actions of some indicated that they had understood or were offering another possibility but there was nothing necessarily to write down."

2.4.4 Forming coherent arguments

Two frequent modes of assessing this skill were evident in the case studies. Written assignments revealed some strengths and weaknesses of argumentation. The other form used involved teacher/student discussions where the teacher provided immediate oral feedback where there was incoherence in the argumentation.

In the *Proof of the pudding CS1*, students advocated their solution to the problem by comparing it to the work of other groups. This task provided the opportunity to revise and improve their own initial solution by collecting input from the whole group.

In the *Chemical reaction speed CS1*, students insisted on their initial erroneous predictions that the relationship between variables was a linear one, but with the help of the teacher they accepted that an exponential function fit the data better. The students worked with the data and realised that thus far they had in fact been working not with the speed of the reaction but with the time needed for the reaction. The rate of reaction was proportional to the reciprocal of the reaction time. This shift from one aspect (variable) to the other was facilitated by data analysis.

In the *Ultraviolet radiation CS1*, a general problem observed was that the students, despite clear instructions from the teacher, did not discuss their inquiry plans with the teacher and often went from questions to investigations without reflection on the planning processes. This led to many experiments with very weak or even false conclusions. However, the discussions seemed to lead the students towards a higher understanding of doing inquiry.

In the *Ultraviolet radiation CS2*, although each student group came to the conclusion that water does not protect from UV radiation, two posters made by two groups indicated very different performance levels in forming coherent arguments. A high performing students' poster reflected the inquiry process through connecting the phases of inquiry into a coherent argument.

The *Martian bacteria CS1* used three-level rubrics in the assessment of this skill. Within this draft unit and case study, for some skills three level, while for others four level rubrics were defined. For forming coherent arguments, the lowest level of the rubric referred to the lack of the skill component, and the middle level indicated either inappropriateness or incompleteness.

In the *Natural selection CS2*, the students should conclude that directional (various types of selection) and random (as genetic drift) processes acting on casual phenotypic variability (conditioned by the genetic one) result in changes in allele frequency in populations, and thus in micro evolutional changes. For the correctly formulated conclusion, 3 points were given if the dependence between natural selection and genetic drift in the course of the evolution was considered. In the *Natural selection CS1*, during the discussion the students were willing to listen to others arguments and carry on the discussion from there. After the exercise, the video recordings were also analysed according to a taxonomy for engagement. There were three categories in this taxonomy going from i) passive, where the students are doing what they are told but nothing more, to ii) participating, where the students take initiative by themselves, towards iii) organizing, where the students also start to organize the group work and assign work to the other group members. For students who showed progression in their conceptual understanding of natural selection (measured through a pre- and post-test), there was a general movement upwards in this taxonomy.

Summary

This report aimed to summarize the experiences of different assessment practices of classroom inquiry. Although some case studies indicated the possibility of summative assessment forms (e.g., when scores from different rubrics were summarized and converted to school marks), most of the practices presented serve the aim of formative and diagnostic assessment.

Assessing content knowledge often results in a diagnosis about students' semantic networks and misconceptions. Mind mapping and brainstorming techniques enabled students to reveal their prior and prerequisite knowledge on different topics. Assessing reasoning and literacy components resulted in diagnoses on misuse of logical or inductive thinking processes. In these cases, formative assessment helped students to reformulate and improve their answers.

The assessment practices on the four highlighted inquiry skills constituted the core of the current report. In the majority of case studies considered, three- or four-level rubrics were used in assessing different skills. Both types of scales allowed for both diagnostic and formative assessment, and both could be used either as the source of immediate feedback during class or as items in post-hoc questionnaires or booklets. The diagnostic value of the three- and four-level rubrics depends on the extent to which they are generalizable. We have seen highly generalizable examples of three- and four level scales in the assessment of all inquiry skills involved in this report. The distinction between the lowest and highest levels in these scales varied according to the inquiry skill assessed and according to the task content. In some cases, variety as a quantitative characteristic served as the descriptor of better attainment, while in other cases the completeness and coherence as qualitative characteristics provided the descriptions for better attainment.

Experiences accumulated so far informed the project that three- and four levels can be readily used in the assessment of skills during classroom discourse. Furthermore, at most two or three inquiry skills may be assessed within one lesson. Taken into account that in many cases group work is an important part of the classroom situations, and since assessment may be realized in both individual and group levels, the strategies of assessing inquiry skills require concentrated efforts by the teachers.

The case studies on the assessment strategies of inquiry skills also inform other work packages. The instruments developed and already used offer the possibility of any comparisons between countries, age groups and inquiry skills. For designing teacher training sessions (WP4), the currently available case studies provide not only usable and nice examples of how to focus on inquiry skills, but draw attention to the importance of lesson planning. The teacher must consciously plan which inquiry skills will be focused on within the lesson and when and how they will be assessed.

The report on the evaluation of the draft units and cases studies is given in D3.2. Further elaboration and clarification of SAILS units is required following the evaluation to streamline the assessment of particular skills within the context of Physics, Chemistry and Biology and also within a more generalizable framework.

Reference

Bybee, R. W. (2009). The BSCS 5E instructional model and 21st century skills. [Available on-line: http://itsisu.concord.org/share/Bybee_21st_Century_Paper.pdf] A commissioned paper prepared for a workshop on exploring the intersection of science education and the development of 21st century skills.

Fradd, S.H., Lee, O., Sutman, F.X., & Saxton, M.K. (2001). Promoting science literacy with English language learners through instructional materials development: A case study. *Billingual Research Journal, 25 (4),* 417-439.

Mullis, I. V., Martin, M. O., Ruddock, G. J., O'Sullivan, C. Y., & Preuschoff, C. (2009). *TIMSS 2011 Assessment Frameworks.* International Association for the Evaluation of Educational Achievement.

OECD (2013). PISA 2012 Assessment and Analytical Framework: Mathematics, Reading, Science, Problem Solving and Financial Literacy, OECD Publishing. http://dx.doi.org/10.1787/9789264190511-en

Wenning, C. J. (2007). Assessing inquiry skills as a component of scientific literacy. *Journal of Physics Teacher Education Online*, *4(2)*, 21-24.