Report from the FP7 project:

Assess Inquiry in Science, Technology and Mathematics Education

ASSISTME

Synthesis of factors determining the uptake of various assessment methods in different educational settings

Delivery date	29/06/2016
Deliverable number	5.11
Lead participant	University of Cyprus
Contact person	C. P. Constantinou and N. Papadouris
Dissemination level	PU

Contents

1. Introduction 4
2. Assessment method: Marking (Grading and Written Comments) 6
2.1 Information on how written feedback was incorporated in the teaching intervention . 6
2.2 Research Data 11
2.3 Data analysis 11
2.4 Results 13
3. Assessment method: Self and Peer Feedback 25
3.1 Information on how peer feedback was incorporated in the teaching intervention ... 25
3.2 Research Data 30
3.3 Data analysis 30
3.4 Results 33
3.4.1 Analysis of feedback comments provided by students 33
3.4.2 Assessors' actions 46
3.4.3 Assessees' actions 47
4. Assessment method: On the fly Interactions 51
4.1. Information on how on the fly Interactions was incorporated in the teaching intervention 51
4.2 Research Data 57
4.3 Data Analysis 58
4.4 Results 60
5. Assessment method: Structured Assessment Dialogue 74
5.1 Information on how structured assessment dialogue was incorporated in the teaching intervention 74
5.2 Research Data 75
5.3 Data analysis 76
5.4 Results 76
5.4.1 Analysis of teacher-students dialogues 76
5.4.2 Analysis of student self-reflections 84
6. Teacher Self-Reporting tools 87
6.1 Written feedback method 87
6.2 Peer feedback method 90
6.3 On the fly assessment method 93
6.4 Structured assessment dialogue 97
6.5 Synopsis 100
7. Synthesis of key findings from classroom trials 101
References 104

Appendix 105

1. Introduction

The ASSISTME project investigated how STEM teachers from each of the eight partner countries implemented and enacted classroom assessment. While clearly the teachers already had assessment practices in place, each partner elected to focus on specific assessment methods to help the project begin to understand how these worked within the inquiry context. Work package (WP) 5 comprised of the trial implementation of the four assessment methods that were selected by the consortium (i.e., Marking (Grading and Written comments), Self and Peer feedback, On the fly interactions and Structured assessment dialogue), by embedding them in teaching/learning materials for certain competences, which were also identified and selected at the level of the consortium.

This report presents the results from all three rounds of implementations. It has emerged from the synthesis of the information that was provided by the Local Working Groups (LWGs) in the corresponding partner countries. It is structured into five sections. The first provides an overview of the scope of the implementations that were undertaken in all three rounds, at the collective level. This provides a synopsis of the context(s) (educational level, subject, competence under emphasis) that were attached to the implementation of each assessment method by the various LWGs in the participating countries. Each of the next four sections focuses on one of the four assessment methods under emphasis in the project. Thereafter, we present an overview of teachers' self-reported perceptions about the assessment methods they enacted. Finally, the last section presents a synopsis of the conclusions derived from the data analysis.

The report synthesizes information from a total of 144 implementations reported by the partners. 33 implementations focused on the assessment method *marking (grading and written comments)*, 34 were based on *self/peer feedback*, 40 drew on *interactions on the fly* and, finally, 37 implementations focused on *structured assessment dialogue*. These implementations provided a broad coverage within the consortium in terms of the educational levels, the subjects and the competences involved in the teaching materials that embodied the assessment methods. Table 1 provides an overview of the number of implementations associated with each assessment method and the corresponding number of teachers and students who participated in these implementations.

Table 1

Implementations (all three rounds of iteration)

Assessment method	Marking (Grading and Written Comments)	Self and Peer Feedback	Interactions on the fly	Structured assessment dialogue
Implementations	33	34	40	37
Total number of students	798	1008	973	approx. 750
Total number of teachers	29	32	24	15

The next four sections of the report summarize key insights reported by the LWGs with respect to the field testing of each of the four assessment methods. The implementation of each assessment method was guided by a specific research design. There was a separate research design for each assessment method, i.e., a total of four different research designs, which are provided in Appendix IV. These research designs were developed in the context of WP5, as a means to ensure sufficient uniformity across the trials with respect to (a) targeted research questions, (b) type of research data that were collected and (c) how these data were processed.

2. Assessment method: Marking (Grading and Written Comments)

2.1 Information on how written feedback was incorporated in the teaching intervention

The implementation of this assessment method conformed to certain specifications, which were formulated at the level of the project with the intent to ensure a robust structure that could allow studying the features of written feedback, as a formative assessment method, in operation. These specifications were as follows:

- The LWGs developed teaching and learning materials targeting one of the six competences under investigation.
- As part of the teaching and learning materials, students submitted to the teacher certain artefacts they had produced, associated with the competence/sub-competence under emphasis. These artefacts could be produced either by individual students or by groups of students.
- The teacher undertook to provide written feedback to each student (or each group of students,). For this purpose, the LWGs sought to develop and provide the teacher with tools intended to facilitate his/her attempt to diagnose students' needs or difficulties, with respect to the competence/sub-competence under emphasis, but also their achievements. To this end, LWGs formulated lists of key aspects of the competence. The depth and breadth afforded by the key aspects were formulated to be consistent with the characteristics of the corresponding target student population.

In total, during the whole project, this specific assessment method was implemented in 33 cases. Tables 2, 3 and 4 summarize the parameters associated with each implementation, in terms of the targeted competence, educational level, number of participating students, and subject and specific topic in which the teaching materials were situated.

With respect to competence, the majority of the implementations were situated in the context of the competence of investigation (26). The remaining nine implementations were associated with argumentation (five), engineering design in technology (two) and modelling and investigation (two). Regarding educational level, the implementations focused mostly on secondary schools. Specifically, twelve of the 33 implementations were implemented in lower secondary, seventeen were tested in upper secondary grades and the remaining four in primary education. The implementations covered different domains of Natural Sciences (8 in Physics, 7 in Biology, 5 in Chemistry,9 in Integrated Science and Technology (4) (Figure 1).

Table 2

Implementations based on marking (grading and written comments) in the first round

	Competence	Educational level	Number of students	Subject	Topic	Country
1	Investigation	Lower secondary	29	Chemistry	Metals	Germany
2	Investigation	Upper secondary	11	Physics	Bungee Jumping	Germany
3	Investigation	Upper secondary	14	Biology	Physical Activity	Denmark
4	Investigation/ modelling	Upper secondary	30	Technology	Electronics	Denmark
5	Investigation/ Modelling	Upper secondary	26	Technology	Electronics	Denmark
6	Investigation	Upper secondary	28	Biology	Health and diet	Denmark
7	Investigation	Lower secondary	35	Biology	Indoor climate	Denmark
8	Argumentation	Primary	17	Integrated Sciences	Flamingo in Larnaca's Salt Lake	Cyprus
9	Argumentation	Primary	18	Integrated Sciences	Flamingo in Larnaca's Salt Lake	Cyprus
10	Investigation	Upper secondary	12	Physics	Kinematics-Free fall	Cyprus
11	Engineering design in Technology	Lower secondary	20	Technology	Materials-Mobile phone stand	Cyprus
12	Investigation	Primary	16	Integrated Science	Human body	Switzerland

Table 3

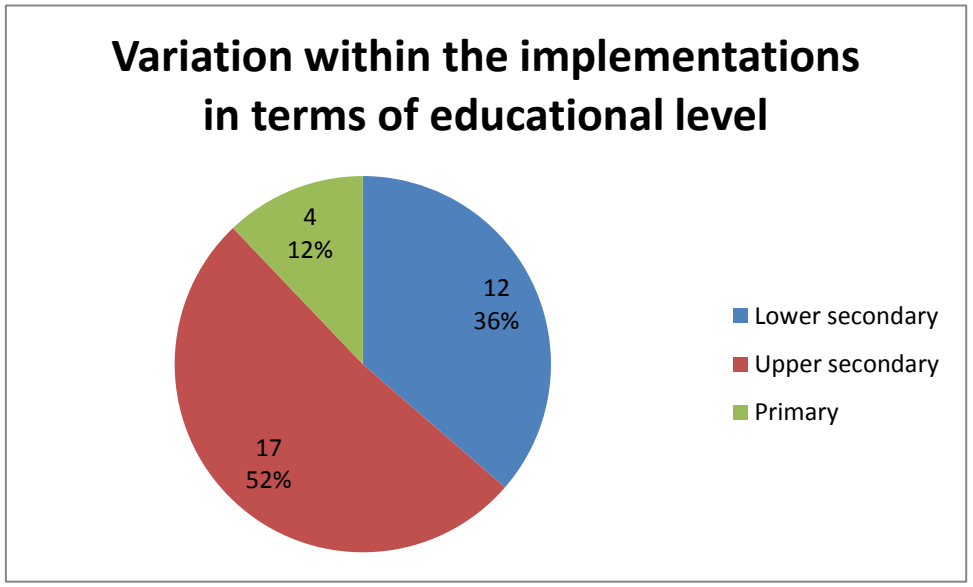
Implementations based on marking (grading and written comments) in the second round

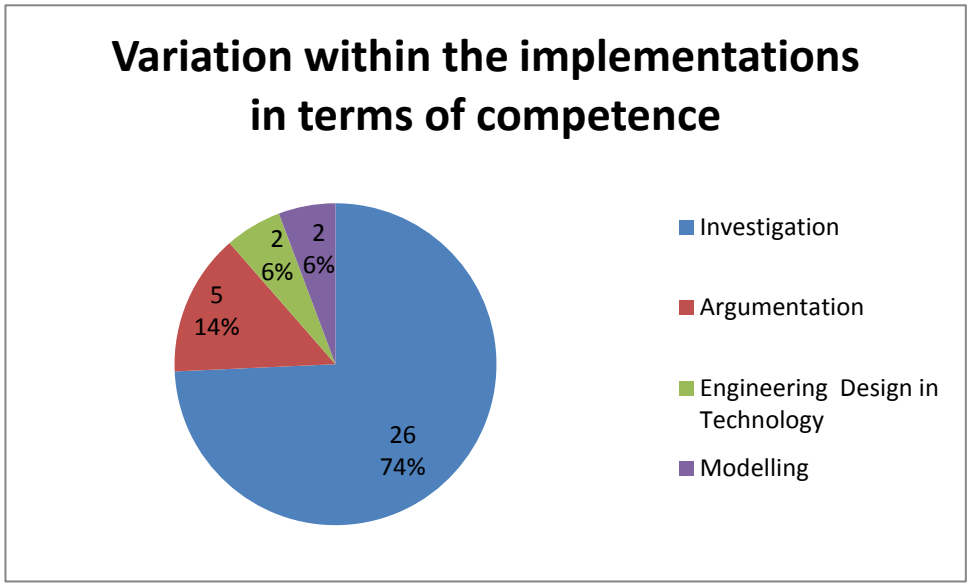
	Competence	Educational level	Number of students	Subject	Topic	Country
1	Investigation	Lower secondary	22	Chemistry	Water (salts)	Germany
2	Investigation	Upper secondary	16	Chemistry	Washing detergents	Germany
3	Investigation	Upper secondary	20	Biology	Aquatic ecology	Switzerland
4	Investigation	Primary	21	Integrated Science	Growing of chicks (=baby hens)	Switzerland
5	Argumentation	Lower Secondary	23	Integrated Sciences	Flamingo in Larnaca's Salt Lake	Cyprus
6	Investigation	Upper secondary	18	Physics	Kinematic-Free Fall	Cyprus
7	Investigation	Upper secondary	18	Physics	Motion on a ramp	Cyprus
8	Engineering Design in Technology	Lower secondary	29	Technology	Sport emblems	Cyprus

Table 4

Implementations based on marking (grading and written comments*) in the third round*

	Competence	Educational level	Number of students	Subject	Topic	Country
1	Investigation	Upper secondary	21	Physics	Electric circuits	Switzerland
2	Investigation	Upper secondary	23	Physics	Science in the city	Switzerland
3	Investigation	Upper secondary	19	Biology	Blood sugar regulation	Denmark
4	Investigation	Lower secondary	17	Integrated science	Human nutrition	Denmark
5	Investigation	Lower secondary	19	Integrated science	Human nutrition	Denmark
6	Investigation	Upper secondary	28	Biology	Physical fitness rating	Denmark
7	Investigation	Upper secondary	22	Biology	Blood sugar regulation	Denmark
8	Investigation	Lower secondary	24	Chemistry	Salts	Germany
9	Investigation	Lower secondary	19	Chemistry	Metals	Germany
10	Argumentation	Lower Secondary	20	Integrated Sciences	Flamingo in Larnaca's Salt Lake	Cyprus
11	Argumentation	Lower Secondary	20	Integrated Sciences	Flamingo in Larnaca's Salt Lake	Cyprus
12	Investigation	Upper secondary	17	Physics	Kinematic-Free Fall, Spring elongation	Cyprus
13	Investigation	Upper secondary	18	Physics	Kinematic-Free Fall, Spring elongation	Cyprus





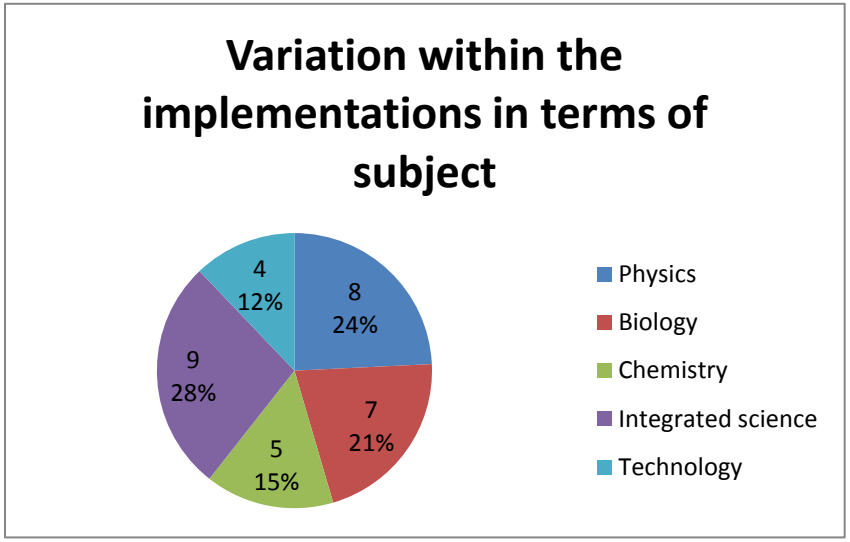


Figure 1. Parameters associated with the implementations of all three rounds

2.2 Research Data

In accordance with the specifications that were formulated for the implementations of this assessment method, the students produced certain artefacts, associated with the competence addressed by the teaching/learning materials that were used, and the teacher produced written feedback comments for these artefacts. The feedback comments prepared by the teacher (in combination with the artefact that was produced by the students) served as the primary data source. An additional data source was the teachers' self-reported views about the assessment method combined with data from the interviews with the teacher. Figure 2 illustrates the research method and shows the research data. The detailed research design (Appendix IV) provides further information about a) the focus of the research design, b) the specific research questions associated with the assessment method, c) the rationale, and d) the constraints that had to be satisfied.

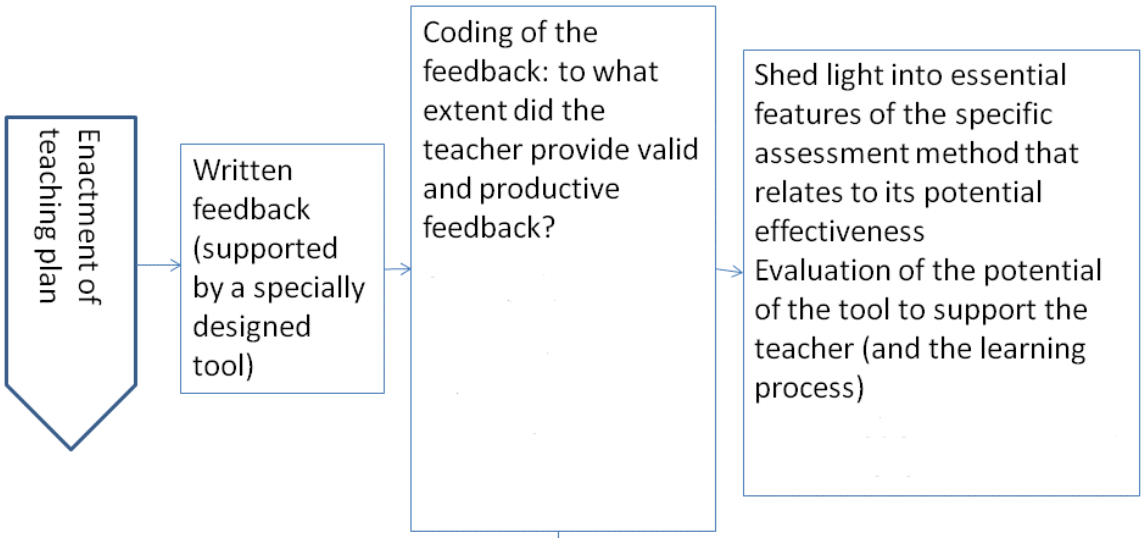


Figure 2. Illustration of the Research Design

2.3 Data analysis

The researchers in the respective countries undertook to process and code the data using a specially designed template (Appendix I). All the completed templates were then collected and synthesized with the intent to detect and document emerging patterns about aspects of the teachers' feedback comments.

The template revolves around four components, described next:

- *Alignment of feedback comments with the targeted aspects of the competence under investigation:* As mentioned earlier, the LWGs were expected to delineate the scope of the teaching/learning materials in terms of the breadth and depth of coverage of the targeted competence. This included the identification of specific learning objectives, in accordance with the characteristics of the target student population. This in turn, yielded a limited number of competence-related criteria against which the artefacts were to be assessed by the teacher. These criteria were pre-specified by the LWGs. The first feature of the feedback comments that was coded refers to the extent to which the teachers committed to this set of criteria and provided feedback comments that did address them.

- *Provision of sufficient justification to the students about the assessment:* The second feature of written feedback that was coded relates to the extent to which the teacher provided sufficient justification to the students about both (a) what they seemed to have achieved with respect to the targeted competence - and were provided credit for – and (b) what they have yet to achieve (Hattie & Timperley, 2007).

- *Provision of guidance:* The third feature relates to the guidance that was provided to the students as to how to move from where they are at, towards the targeted learning goals associated with the competence of interest (Hattie & Timperley, 2007). The teachers' feedback comments were coded on the basis of whether they included guidance about specific, concrete next steps.

- *Consideration of affective dimensions:* The last feature that was taken into account in coding teacher's feedback relates to the extent to which they addressed affective dimensions, which are considered to be of paramount importance in terms of how the students experience the process of engaging with feedback (Värlander, 2008).

Each of the aspects mentioned above was coded on an ordinal (Likert) scale, which was embedded in the template. Table 5 shows the range of the scale for each of the four aspects and illustrates how it was used to capture the corresponding variation across the feedback comments.

The data for the various sections of the completed templates from the various LWGs were aggregated in a single file and were then exposed to analysis, using descriptive statistics, which was supplemented, where useful, with appropriate non-parametric tests (e.g. Wilcoxon signed-rank test).

The teacher's self-reporting tool consisted of a set of fourteen Likert scale items and four open-ended questions about the teacher's perceptions concerning (a) the process of enacting the specific assessment method in their context (e.g. time requirements), (b) how certain features of the assessment method played out, and (c) how students engaged with the assessment method (Appendix I). The data from this source were processed through descriptive statistics. The results of the analysis are presented in section 6.

Table 5

Features of the feedback comments that were coded

Component	Likert-scale	Description
Number of criteria	1-5	1=The feedback comments did not take into account any of the relevant aspects 5=The feedback comments took into account all relevant aspects
Provision of justifications		
Justification on what has been achieved	1-5	1=None of the references to what was achieved was accompanied with a relevant justification 5= All references were justified
Justification on what has yet to be achieved	1-5	1= None of the references to what was not achieved, was accompanied with a relevant justification 5= All references were justified
Provision of guidance	1-4	1= The teacher did not provide specific guidance to students 4=The teacher provided specific guidance about concrete next steps
Affective Dimension		
Number of affective elements evident in the feedback comments	1-4	1=The feedback comments did not include affective elements 4=The feedback comments included many affective comments
Direction of the affective side of the comments (positive or negative)	1-5	1=Overall the affective elements were mostly negative 5=Overall the affective elements were mostly positive

2.4 Results

This section describes the results of the analysis of teachers' feedback comments that emerged from the synthesis of the data reported from the three rounds of implementations. The analysis of each component is presented separately.

Analysis of feedback comments provided by the teachers

Extent to which the teacher's feedback comments took into account the relevant aspects of the competence

Overall, the feedback comments of all the rounds appeared to be rather thorough in terms of coverage of the relevant criteria that were specified by the teachers and discussed with the students. These criteria were reflecting the key aspects of the competence of interest. The mean value for this variable

and the median indicate that in their majority the teachers offered comprehensive comments that addressed most of the specified criteria ($\bar{x}$=4.42, Mdn=5, SD=0.783). Tables 6, 7 and 8 report results from descriptive statistics about this aspect. These are shown separately for each country per round. Also, Figure 3 provides an example of how a teacher incorporated all the intended criteria of the investigation competence.

Table 6

Number of criteria considered by the teachers in the first round

First round						
Country	N	Observed minimum	Observed maximum	Median	Mean	Standard Deviation
Germany	29[1]	2	5	4	3.90	0.673
Denmark	53	1	5	4	3.51	0.993
Cyprus	48	3	5	5	4.52	0.583
Total	**130**	**1**	**5**	**4**	**3.97**	**0.906**

Table 7

Number of criteria considered by the teachers in the second round

Second round						
Country	N	Observed minimum	Observed maximum	Median	Mean	Standard Deviation
Germany	36	3	5	5	4.94	0.333
Switzerland	38	4	5	5	4.76	0.431
Cyprus	67	2	5	4	4.22	0.885
Total	**141**	**2**	**5**	**5**	**4.55**	**0.741**

Table 8

Number of criteria considered by the teachers in the third round

Third round						
Country	N	Observed minimum	Observed maximum	Median	Mean	Standard Deviation
Germany	24	5	5	5	5	0.000
Switzerland	21	2	5	5	4.67	0.730
Denmark	71	4	5	5	4.68	0.471
Cyprus	57	2	5	5	4.44	0.598
Total	**173**	**2**	**5**	**5**	**4.64**	**0.549**

[1] N refers to the number of artefacts on which the teacher offered feedback

Criteria:

a) Specify a procedure that would allow changing the independent variable in a systematic manner that could allow exploring whether it does influence the dependent variable
b) Specify a procedure that would allow accurate and sufficiently sensitive measurements of possible changes in the dependent variable as a result of the (purposeful) changes in the independent variable
c) Specify a procedure that would properly control for the possible impact of other relevant variables thereby enabling the attributions of changes in the dependent variable to the independent variable

Teacher's feedback comments:

a) You correctly mentioned that you are going to change the mass. You didn't specify a procedure that would allow changing the independent variable.

b) Bravo! You correctly recognized the dependent variable. The way you are going to measure the dependent variable is not clear and should be clarified.

c) The variable you will be changing at your experiment is the mass. Which other variables need to be stable? Describe how for each one. The height is one of them!

Figure 3. An example of how a teacher incorporated all the intended criteria of the investigation competence

A possible interpretation for this finding pertains to the benefits emerging from the establishment and operation of LWGs. In particular, this may have provided a productive structure that served to support teachers in enacting this assessment method. For example, one part of this procedure involved the articulation of the important aspects of the given competence in each case and the discussion of how these were to be negotiated with the students in the class. Another interpretation relates to the use of specially designed tools, intended to systematize teachers' effort to address the relevant aspects of the competence of interest. An example of such tool is given in Appendix I. Such tools could have helped maintaining a high degree of alignment between the focus of the feedback and the essential aspects of the competence of interest.

Provision of justification to the students about the assessment

The analysis of the data in view of the second component, namely the provision of justification to the students about the assessment, involved two stages. Firstly, we examined the references made by the teachers to what the students had achieved, or had not achieved, with respect to the specified criteria. As shown in Table 9, references to what the students had achieved appeared in 87% whereas references to what they had yet to achieve occurred in 70% of the cases.

Table 9

References about student's level of attainment in the three rounds

		First Round	Second Round	Third Round	**Total**
What the student has achieved	No references	8	19	8	**13**
	References	92	81	92	**87**
What the student has yet to achieve	No references	35	31	24	**30**
	References	65	69	76	**70**

In the second stage, we distinguished those references into two categories depending on whether they were accompanied by a relevant justification addressed to the student. A reference was considered justified in cases when the teacher provided reasons to the students, as to how s/he knew that the student had, or had not, achieved a certain criterion (or, put differently, had, or had not, attained a specific aspect of the competence of interest). Figure 4, shows a translated example of a justified reference to what the student had not achieved based on the initial artefact she had produced. The justified references to what the students had not achieved ($\bar{x}$=3.20, Mdn=3, SD=1.149) were more frequent than the justified references to what they had achieved ($\bar{x}$=1.79, Mdn=1, SD=1.408). Tables 10 – 15 report descriptive statistics about this aspect of the teachers' feedback comments for each round. Interestingly, even though the teachers referred to what the student has achieved more frequently (87%) than to what they had yet to achieve (70%), they happened to provide more justified references about students' weaknesses ($\bar{x}$=3.20, Mdn=3, SD=1.149) than students' strengths ($\bar{x}$=1.79, Mdn=1, SD=1.408).

One possible reason for this seeming tendency of the teachers to offer more justifications for their references to what their students had not yet accomplished, could be related to their belief that the students could benefit more by those justifications since they might be providing students with an indication as to what they need to improve. This finding is supported by the interview data, as shown in the following excerpt from an interview with a teacher in Cyprus:

> *Interviewer: In most of the cases you provided justification only for what the students had yet to achieve. Why didn't you give such justified references for students' achievements?*
>
> *Teacher: I believe that it is more useful to give justification for students' weaknesses in order to specify the aspects that they have to improve. On the contrary, they do not need any clarification about what they had achieved because it will not actually help them improve.*

Competence: investigation

Topic: Free fall

Investigable question: Does the mass of an object influence the incidence velocity when the object falls from a height?

Type of artefact: description of an experiment, with appropriate control of variables, that would allow addressing the investigable question

Assessment Criteria: (a) identification of the relevant confounding variables, (b) appropriate control of these variables so that changes in the dependent variable can be attributed to corresponding changes in the independent variable, (c) systematic approach of varying the independent variable (d) appropriate approach for measuring changes in the dependent variable.

Student's initial artefact: We will try to keep the height stable. The shape will be kept stable because we are going to fill the box.

Teacher's feedback comments: You didn't mention how you will fill the box in order to keep stable the shape.

Figure 4. Example of a justified reference about what the student has yet to achieve

Table 10

Justification on what has been achieved provided by teachers in the first round[2]

		First Round				
Country	N	Observed minimum	Observed maximum	Median	Mean	Standard Deviation
Germany	29	1	4	2	1.72	0.635
Cyprus	42	1	5	1	1.60	1.369
Total	**71**	**1**	**5**	**1**	**1.65**	**1.060**

[2] In Denmark it was not possible to reach satisfactory interrater reliability on this item and thus their data is not included.

Table 11

Justification on what has been achieved provided by teachers in the second round

Second Round						
Country	N	Observed minimum	Observed maximum	Median	Mean	Standard Deviation
Switzerland	22	1	2	2	1.59	0.253
Germany	38	1	4	3	2.82	1.668
Cyprus	54	1	5	1	1.65	1.251
Total	**114**	**1**	**5**	**1**	**2.03**	**1.495**

Table 12

Justification on what has been achieved provided by teachers in the third round

Third Round						
Country	N	Observed minimum	Observed maximum	Median	Mean	Standard Deviation
Germany	18	1	3	1	1.11	0.471
Switzerland	21	1	1	1	1	0.000
Cyprus	55	1	5	1	1.85	1.311
Total	**94**	**1**	**5**	**1**	**1.52**	**1.095**

Table 13

Justification on what has not been achieved provided by teachers in the first round

First Round						
Country	N	Observed minimum	Observed maximum	Median	Mean	Standard Deviation
Germany	28	1	5	4	4.07	1.274
Cyprus	22	1	5	2.50	2.82	1.842
Total	**50**	**1**	**5**	**4**	**3.52**	**1.657**

Table 14

Justification on what has not been achieved provided by teachers in the second round

Second Round						
Country	N	Observed minimum	Observed maximum	Median	Mean	Standard Deviation
Switzerland	36	1	4	2	2.33	0.676
Germany	38	1	5	4	3.71	1.160
Cyprus	23	1	5	4	3.48	1.729
Total	**97**	**1**	**5**	**3**	**3.14**	**1.331**

Table 15

Justification on what has not been achieved provided by teachers in the third round

Third Round						
Country	N	Observed minimum	Observed maximum	Median	Mean	Standard Deviation
Germany	24	1	5	3	3.08	1.060
Switzerland	21	2	5	4	3.38	1.396
Cyprus	32	1	5	2	2.59	1.633
Total	**77**	**1**	**5**	**3**	**2.96**	**1.446**

Provision of guidance

Another aspect of the feedback that was coded, by means of the Likert scale items included in the coding tool, focuses on the guidance provided by the teacher to their students as to how to go about improving their artefact and, hence, strengthening underdeveloped aspects of the competence. The mean score overall (for all three rounds) that was assigned to the teachers' feedback comments in terms of the specificity of the guidance they incorporated was 2.70 (out of a maximum value of 4), with a median value of 3 and a standard deviation of 0.781. Tables 16, 17 and 18 show in detail the score for specificity of the guidance provided by the teachers in each round of implementation. Figure 5 presents an example of a feedback comment where the guidance received a high score in terms of its specificity (category 3).

Table 16

Guidance provided by the teachers in the first round

First Round						
Country	N	Observed minimum	Observed maximum	Median	Mean	Standard Deviation
Germany	29	1	4	3	2.66	0.614
Denmark	53	1	4	2	2.62	0.765
Cyprus	48	1	4	2	2.23	0.722
Total	**130**	**1**	**4**	**2**	**2.48**	**0.739**

Table 17

Guidance provided by the teachers in the second round

Second Round						
Country	N	Observed minimum	Observed maximum	Median	Mean	Standard Deviation
Switzerland	34	1	4	3	2.94	0.547
Germany	35	3	4	4	3.77	0.426
Cyprus	67	1	4	2	2.09	0.484
Total	**136**	**1**	**4**	**3**	**2.74**	**0.854**

Table 18

Guidance provided by the teachers in the third round

Third Round						
Country	N	Observed minimum	Observed maximum	Median	Mean	Standard Deviation
Germany	24	3	4	4	3.67	0.482
Switzerland	19	3	4	3	3.68	0.478
Denmark	71	1	4	2	2.44	0.579
Cyprus	57	2	3	3	3	0.462
Total	**171**	**1**	**4**	**3**	**2.84**	**0.717**

Competence: investigation

Investigable question: Does the height that we release an object influence the velocity at which the object reaches the ground?

Criterion: Specify a procedure that would properly control for the possible impact of other relevant variables thereby enabling the attributions of changes in the dependent variable to the independent variable

Student's initial artefact: We will keep constant all variables apart the one we are going to measure.

Teacher's guidance: The variable that you have to keep constant is the type of the object's material (mass and volume of the object). How will you make sure that it will be indeed kept constant?

Figure 5. Example of a guidance

One point that warrants mention relates to the connection between the analysis for the first and the third aspect of the template (i.e., extent to which the teachers' feedback comments took into account the relevant aspects of the competence and provision of guidance). In an attempt to investigate if there is a statistically significant difference between these two variables we employed Wilcoxon signed-rank test. The results seem to be suggesting that teacher's feedback was coded consistently higher on the first variable (extent to which their feedback comments took into account the relevant aspects of the competence) compared to the second (provision of guidance). In view of the relatively high mean, associated with the first variable this finding seems to be suggesting that whereas the teachers tended to provide feedback comments that took into account the relevant criteria, they found it rather difficult to offer specific guidance on concrete next steps ($p < .001$, $z = -15.277$, $r = -.76$). This result should be treated with caution due to the limited sample size. This limitation notwithstanding, we take the perspective that this finding is hardly surprising. In particular, the provision of guidance can be much more creative, but also demanding, in the sense that it deals with aspects that cannot be anticipated beforehand and need to be judged by the individual teachers in a case-by-case manner, depending on their judgment of the content of the artefact and the personal characteristics of the corresponding students.

In any case, one claim that could be made in view of these results, is that there seems to be a need to support teachers in performing this task more effectively. In an attempt to address this, Cyprus in the second implementation sought to develop a specially designed tool which contained a large sample of illustrative examples on which the teacher could draw upon in producing their own written feedback comments (Appendix I). The idea behind this tool was that it could support teachers' attempt to formulate useful feedback. Nevertheless, this hypothesis was not confirmed by the data that emerged in that particular case. Specifically, in many cases (in 34 out of the 46 cases, 74%), the teacher happened to offer feedback comments that were not well aligned with the students' needs (Table 19). An extreme example that we encountered in the data reported by the Cyprus LWG, which could serve to illustrate this unproductive use of the tool involves asking students to include additional features in his/her artefact even though these were already there. One possible interpretation for this seemingly unproductive use of the tool is that the teacher found it convenient to rely on the given paradigmatic examples of feedback that were incorporated in the tool, often not in a thoughtful manner. The key idea that emerged from this is that whereas teachers can benefit substantially from tools intended to scaffold the provision of feedback comments, this should not be taken for granted. For instance, it could be argued that the given examples seemed to have subtracted from the teacher's responsibility to interpret students' products, diagnose needs and reflect on possible useful feedback. Hence, exposing teachers to examples of feedback doesn't always guarantee that teachers will effectively use them to promote their students' learning outcomes. In the concluding part of the report we revisit this claim and we elaborate on its implications.

Table 19

Guidance provided by the teacher before and after the given examples of feedback

	Without the tool		With the tool	
	Aligned with the student's artefact	Not aligned with the student's artefact	Aligned with the student's artefact	Not aligned with the student's artefact
Frequency	29 (91%)	3 (9%)	12 (26%)	34 (74%)
Total	**32 (100%)**		**46 (100%)**	

Affective dimension

The last component that was addressed in the analysis of the teachers' feedback relates to their affective side. For this we examined the number of the elements of the feedback comments that carried affective connotations (e.g. "Bravo! You correctly identified the independent variable and specified a procedure that allows changing the independent variable") and we then investigated whether these connotations were mostly positive or negative. It could be argued that the teachers did not appear to be very attentive to the affective side since their feedback received a relatively low score, overall, on this dimension ($\bar{x}$=1.81, Mdn=2, SD=0.802). Tables 20, 21 and 22 indicate the number of affective comments provided by the teachers in each round of implementation. It is also important to note that in the cases where the feedback comments did contain affective connotations, these were overall rather balanced with a tendency towards the positive side ($\bar{x}$=3.38, Mdn=4, SD=1.198). Tables 23, 24 and 25 report descriptive statistics about the affective aspect of the feedback comments, in each of the three rounds.

Table 20

Number of affective comments provided by the teachers in the first round

First Round						
Country	N	Observed minimum	Observed maximum	Median	Mean	Standard Deviation
Germany	29	1	4	2	1.62	0.677
Denmark	53	1	3	2	1.79	0.495
Cyprus	47	2	4	2	2.47	0.678
Total	**129**	**1**	**4**	**2**	**2**	**0.707**

Table 21

Number of affective comments provided by the teachers in the second round

Second Round						
Country	N	Observed minimum	Observed maximum	Median	Mean	Standard Deviation
Switzerland	36	1	4	2	1.89	0.919
Germany	38	1	3	1	1.42	0.552
Cyprus	67	1	4	2	2.13	0.776
Total	**141**	**1**	**4**	**2**	**1.88**	**0.815**

Table 22

Number of affective comments provided by the teachers in the third round

Third Round						
Country	N	Observed minimum	Observed maximum	Median	Mean	Standard Deviation
Germany	24	1	2	1	1.13	0.338
Switzerland	21	1	1	1	1	0.000
Denmark	71	1	4	1	1.49	0.694
Cyprus	57	1	4	2	2.19	0.895
Total	**173**	**1**	**4**	**1**	**1.61**	**0.818**

Table 23

Potential affective impact (positive or negative) on students (first round)

First Round						
Country	N	Observed minimum	Observed maximum	Median	Mean	Standard Deviation
Germany	16	2	5	4	3.31	1.138
Denmark	40	1	4	3	2.88	0.723
Cyprus	47	2	5	5	4.47	0.830
Total	**103**	**1**	**5**	**4**	**3.67**	**1.124**

Table 24

Potential affective impact (positive or negative) on students (second round)

Second Round						
Country	N	Observed minimum	Observed maximum	Median	Mean	Standard Deviation
Switzerland	36	1	3	2	1.78	0.760
Germany	15	2	4	2	2.93	1.033
Cyprus	56	3	5	4	4	0.426
Total	**107**	**1**	**5**	**4**	**3.10**	**1.205**

Table 25

Potential affective impact (positive or negative) to students' state in the third round

Third Round						
Country	N	Observed minimum	Observed maximum	Median	Mean	Standard Deviation
Germany	21	2	5	5	4.24	0.944
Switzerland	21	1	2	2	1.52	0.512
Denmark	59	2	5	3	2.95	0.600
Cyprus	47	2	5	4	4.36	0.735
Total	**148**	**1**	**5**	**3**	**3.38**	**1.203**

3. Assessment method: Self and Peer Feedback

3.1 Information on how peer feedback was incorporated in the teaching intervention

The implementation of this assessment method, conformed to certain specifications, which were formulated at the level of the project with the intent to ensure a robust structure that could allow studying the features of peer feedback, as a formative assessment method. These specifications were as follows:

- The LWGs had to develop teaching and learning materials targeting one of the six competences under investigation.
- Before the implementation of the assessment method the students should be introduced to the roles of the peer-assessor and the peer-assessee. During the implementation, students alternate between the two roles (peer-assessor and peer assessee).
- As part of the teaching and learning materials, student(s) had to submit to their peer(s) certain artefacts they had produced, associated with the competence/sub-competence under emphasis. These artefacts could be produced either by individual students or by groups of students.
- The student(s) undertook to provide feedback to his/her (their) peer(s). The process of exchanging peer-feedback is supported through specially designed templates, which encompass criteria for assessing the specific artefacts. These tools are developed by the LWGs. The scope of the criteria should be consistent with the characteristics of the target student population.

In total, this specific assessment method was implemented in 34 cases. Tables 26, 27 and 28 summarise the parameters associated with each of these 34 implementations in terms of the targeted competence, educational level, number of participating students, and subject and specific topic in which the teaching materials were situated (in each round).

With respect to competences, fifteen of the 34 implementations were associated with investigation, twelve were associated with modelling, three with problem solving and the remaining four targeted a combination of two competences (argumentation & problem solving, engineering design in technology & investigation, modelling & problem solving). With regard to educational level, seventeen of the 34 implementations were implemented in the upper secondary education, seven in lower secondary education and ten in primary education. The implementations covered a range of domains, including Physics (7), Biology (15), Geosciences (3), Integrated Sciences (3), Mathematics (5) and Technology (1) (Figure 6).

Table 26

Implementations based on the self and peer assessment method in the First Round

	Competence	Educational level	Number of students	Subject	Topic	Country
1	Modelling	Upper secondary	24	Biology	Immune reaction	Switzerland
2	Modelling	Upper secondary	23	Physics	Electrical circuits	Switzerland
3	Investigation	Upper secondary	64	Physics	Periodic phenomena	France (Lyon)
4	Investigation	Upper secondary	54	Geosciences	Fossil energy	France (Lyon)
5	Modelling/ Problem solving	Primary	48	Mathematics	Big Numbers	Czech Republic
6	Investigation	Lower Secondary	25	Biology	Human physiology	Czech Republic
7	Investigation	Primary	23	Integrated science	How the plant grow up	Czech Republic
8	Problem solving	Upper Secondary	28	Mathematics	Trigonometry	Denmark
9	Modelling	Primary	20	Biology	How plants grow up	Cyprus
10	Modelling	Primary	20	Biology	How plants grow up	Cyprus

Table 27

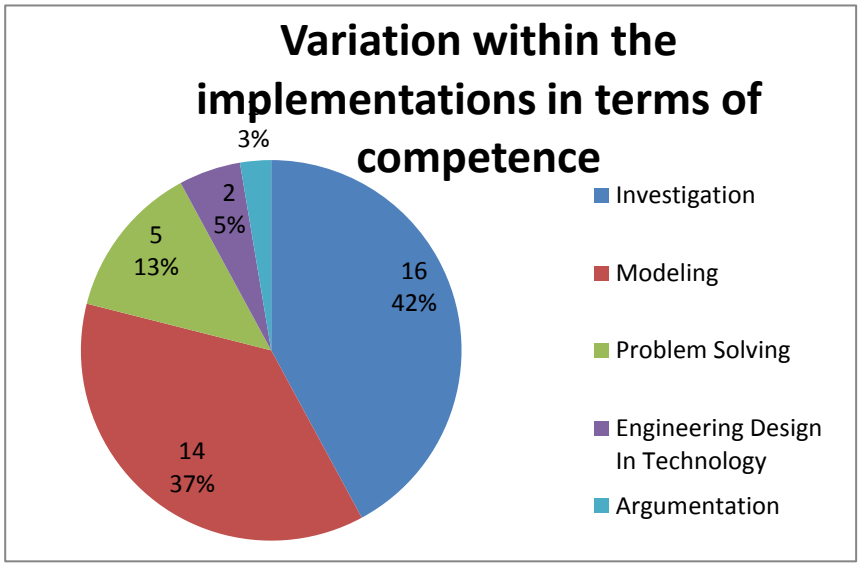
Implementations based on the self and peer assessment method in the Second Round

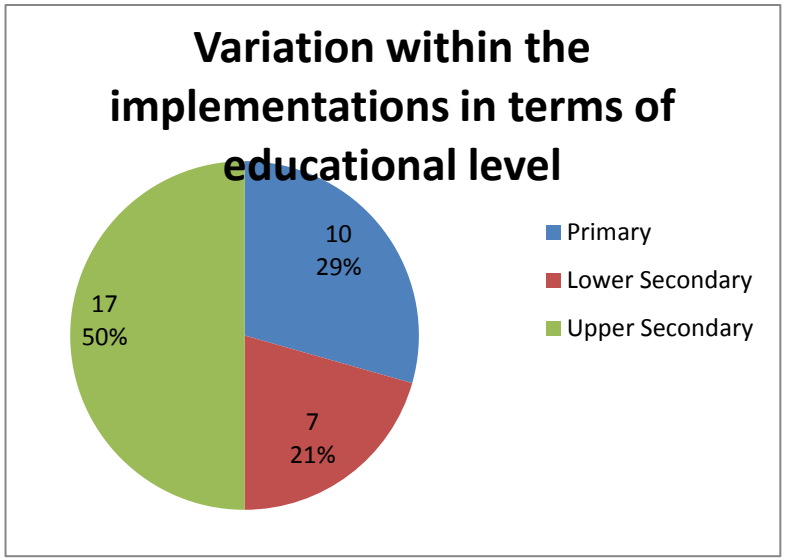
	Competence	Educational level	Number of students	Subject	Topic	Country
1	Modelling	Upper Secondary	22	Biology	Photosynthesis	Switzerland
2	Modelling	Upper Secondary	24	Physics	Optics- Colour and Light	Switzerland
3	Investigation/ Engineering Design In Technology	Upper Secondary	53	Geosciences	Fossil energies	France (Lyon)
4	Investigation	Upper Secondary	55	Physics	Periodic phenomena	France (Lyon)
5	Investigation	Primary	24	Integrated Science	Plant Physiology	Czech Republic
6	Investigation	Lower Secondary	22	Biology	Plant Physiology	Czech Republic
7	Problem Solving/ Modelling	Primary	37	Mathematics	Basic geometrical shapes	Czech Republic
8	Engineering Design In Technology	Upper Secondary	25	Technology	Needs analysis	Denmark
9	Investigation	Upper Secondary	38	Biology	Metabolic rates	Denmark
10	Investigation	Upper Secondary	19	Biology	Genetics and blood type	Denmark
11	Argumentation/ Problem solving	Upper Secondary	21	Mathematics	Programming	Denmark
12	Modelling	Lower Secondary	21	Biology	Various biological cycles	Denmark
13	Modelling	Primary	20	Biology	How plants grow up	Cyprus
14	Modelling	Primary	20	Biology	How plants grow up	Cyprus

Table 28

Implementations based on the self and peer assessment method in the Third Round

	Competence	Educational level	Number of students	Subject	Topic	Country
1	Modelling	Upper secondary	17	Physics	Capacity	Switzerland
2	Investigation	Upper secondary	21	Physics	Electric circuits	Switzerland
3	Investigation	Upper secondary	64	Physics	Periodic phenomena	France (Lyon)
4	Investigation	Upper secondary	59	Geosciences	Fossil energy	France (Lyon)
5	Investigation	Primary	28	Integrated Science	Germination	Czech Republic
6	Investigation	Lower Secondary	12	Biology	Germination	Czech Republic
7	Investigation	Lower Secondary	11	Biology	Breathing frequency	Czech Republic
8	Problem solving	Primary	36	Mathematics	Area and perimeter of geometrical shapes	Czech Republic
9	Modelling	Lower Secondary	20	Biology	How plants grow up	Cyprus
10	Modelling	Lower Secondary	19	Biology	How plants grow up	Cyprus





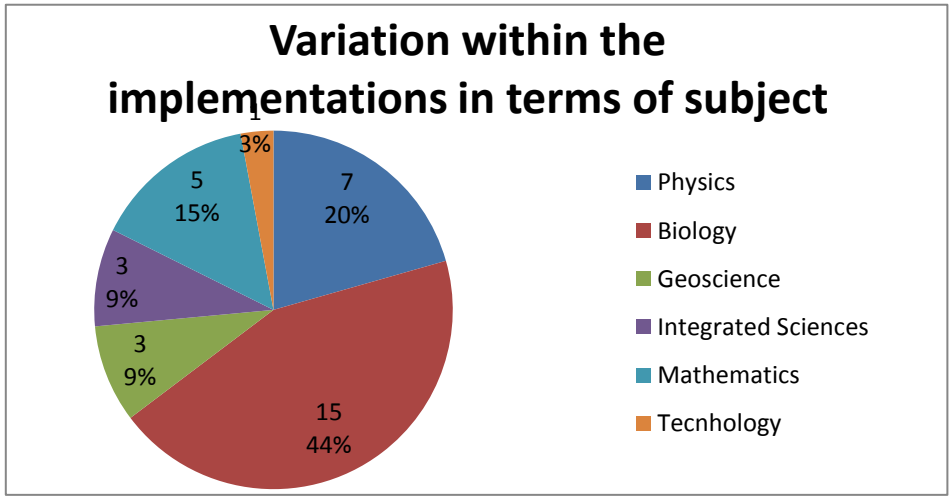


Figure 6. Parameters associated with the implementations of all three rounds

3.2 Research Data

In accordance with the specifications that were formulated for the implementations of this assessment method, the students (as assessees) produced certain artefacts, associated with the competence addressed by the teaching/learning materials. Each student (group of students) then undertook to assess the artefact produced by another student (group) and produced written feedback comments (assessors). The feedback comments prepared by the peer-assessors (in combination with the initial and revised artefacts that were produced by the peer-assessees) served as the primary data source. An additional data source included the data from the interviews with students and the teachers' self-reporting tools. Figure 7 provides an overview of the research method and the research data. The detailed research design (Appendix IV) provides further information about the a) focus of the research design, b) specific research questions associated with the assessment method, c) rationale, and d) constraints to be satisfied.

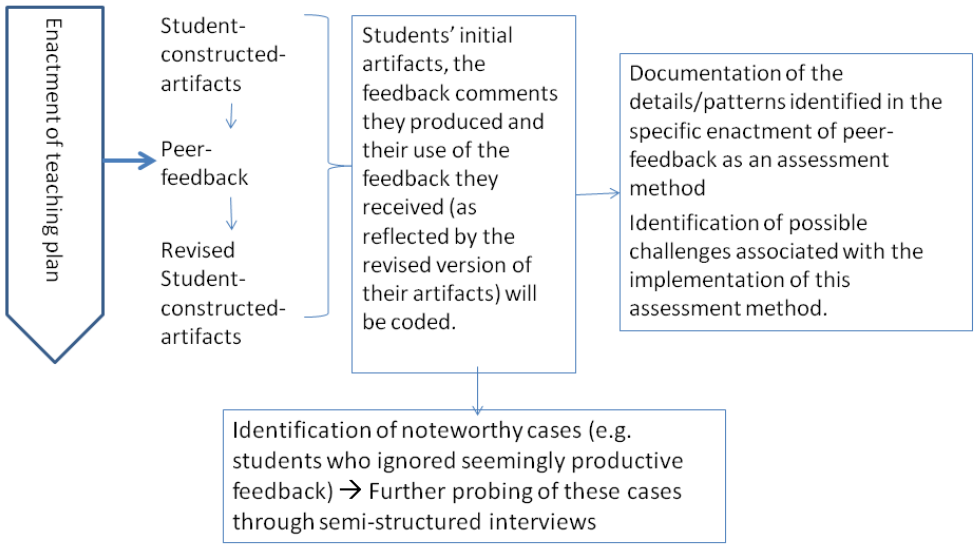


Figure 7. Illustration of the Research Design

3.3 Data analysis

The LWGs undertook to process and code the data using a specially designed template (Appendix II). All completed templates were then collected and synthesized with the intent to detect and document emerging patterns about aspects of the students' feedback comments and peer-assessors'/assessees' actions.

The template revolves around the following five components:

- *Consideration of affective dimensions:* The first feature that was taken into account in coding student's feedback relates to the extent to which they addressed affective dimensions, which

are considered to be of paramount importance in terms of how the students engage with the process of exchanging feedback (Värlander, 2008).

- *Alignment of feedback comments with the targeted aspects of the competence under investigation:* As mentioned earlier, the LWGs were expected to delineate the scope of the teaching/learning materials in terms of the breadth and depth of coverage of the targeted competence. This included the identification of specific learning objectives, in accordance with the characteristics of the target student population. This, in turn, yielded a limited number of competence-related criteria against which the artefacts were to be assessed by the student. These criteria, were pre-specified by the LWGs. The feature of the feedback comments that was coded, refers to the extent to which the teachers committed to this set of criteria and provided feedback comments that did address them (Sadler, 1989; Tsivitanidou et al., 2010).

- *Reliance of the assessor or the criteria while constructing his/her own artefact*
  This relates to the extent to which the peer-assessor(s) drew on the pre-specified criteria determined by the LWG while preparing his/their own artefact.

- *Provision of sufficient justification to the students about the assessment:* The fourth feature of the feedback comments that was coded, relates to the extent to which the peer-assessor provided sufficient justification to the peers about both (i) what they seemed to achieve with respect to the targeted competence - and were provided credit for – and (ii) what they had yet to achieve (Hattie & Timperley, 2007; Gielen et al., 2010).

- *Provision of guidance:* The last feature relates to the guidance that was provided to the peer-assessees as to how to move from where they are at, towards the targeted learning goals associated with the competence of interest (Hattie & Timperley, 2007). In particular, the peers' feedback comments were coded on the basis of whether they included guidance about specific, concrete next steps.

Each of the aspects mentioned above was coded on an ordinal (Likert) scale, which was embedded in the template. Table 29 shows the range of the scale for each of the five aspects and illustrates how it was used to capture the corresponding variation across the feedback comments.

Table 29

Features of the feedback comments that were coded

Component	Likert-scale	Description
Affective dimension		
Direction of the affective side of the comments (positive or negative)	1-5	1=Overall the affective elements were mostly negative 5=Overall the affective elements were mostly positive
Alignment of feedback comments with the targeted aspects of the competence		
Number of the criteria	1-4	1=The feedback comments did not take into account any of the relevant aspects 4=The feedback comments took into account all relevant aspects
Extent to which the peer-assessor draw on the criteria in a thorough manner	1-4	1=The assessor drew on the criteria in a superficial manner 4= The assessor drew on the criteria in a thorough manner
Reliance of the assessor on the criteria while constructing his/her own artefact		
Number of the criteria	1-4	1=The assessor did not draw on any of the criteria 4= The assessor drew on all criteria
Extent to which the peer-assessor draw on the criteria in a thorough manner	1-4	1=The assessor drew on the criteria in a superficial manner 4= The assessor drew on the criteria in a thorough manner
Provision of justification		
Justification on what has achieved	1-5	1=All references were without justification 5= All references were justified
Justification on what has yet to be achieved	1-5	1= All references were without justification 5= All references were justified
Provision of guidance		
Specificity of the guidance	1-4	1= The peer-assessor did not provide specific guidance about concrete next steps 4=The peer-assessor provided specific guidance about concrete next steps
Related to the competence of interest	0-1	0=No 1=Yes

The data for the various sections of the templates that were completed by the individual LWGs were aggregated in a single file and were then exposed to analysis using descriptive statistics.

The teacher's self-reporting tool consisted of a set of eleven Likert scale items and four open-ended questions about the teacher's perceptions concerning (a) the process of enacting the specific assessment method in their context (e.g. time consuming), (b) how certain features of the assessment method played out, and (c) how students engaged with the assessment method (Appendix II). The data from this source were processed through descriptive statistics. The results of the analysis are presented in section 6.

3.4 Results

This section presents the results that emerged from the synthesis of the data from the implementations of peer feedback. It is structured in two sub-sections. The first pertains to the analysis of the feedback comments provided by peer-assessor(s), considering the various sections that comprised the specially designed template. The analysis of each component is presented separately. The second relates to how the peer-assessees or the peer assessors reacted to their corresponding roles.

3.4.1 Analysis of feedback comments provided by students[3]

Affective dimension

The first component that was addressed relates to the affective side of the feedback comments. For this, we examined the number of the elements of the feedback comments that carried affective connotations and we then investigated whether these were mostly positive or negative. One example of positive affective connotations from one LWG from the Czech Republic is the following: "It is explained gorgeously and written nicely as well". Overall, the feedback comments appeared to be balanced ($\bar{x}$=3.22, Mdn=3, SD=0.882). Tables 30, 31 and 32 show the results considering the affective dimension in every country in each round. The available data seem to be suggesting that the students, as peer-assessors, did not pay much attention to the affective dimension of their comments.

Table 30

[3] Some of the items of the coding tool were optional, thus some countries do not appear on some of the tables.

Affective impact (positive or negative) in the first round

First Round						
Country	N	Observed minimum	Observed maximum	Median	Mean	Standard Deviation
Cyprus	13	1	5	3.00	3.46	1.664
Switzerland	44	1	5	3.00	3.36	1.313
Czech Republic	62	1	5	3.00	3.40	0.858
Denmark	5	2	4	3.00	3.00	0.707
Total	**124**	**1**	**5**	**3.00**	**3.38**	**1.123**

Table 31

Affective impact (positive or negative) in the second round

Second Round						
Country	N	Observed minimum	Observed maximum	Median	Mean	Standard Deviation
Cyprus	8	3	5	3.00	3.00	0.000
Switzerland	45	1	5	4.00	3.47	1.392
Czech Republic	217	1	5	3.00	3.08	0.563
Denmark	51	1	5	3.00	3.08	0.771
Total	**321**	**1**	**5**	**3.00**	**3.13**	**0.769**

Table 32

Affective impact (positive or negative) in the third round

Third Round						
Country	N	Observed minimum	Observed maximum	Median	Mean	Standard Deviation
Cyprus	11	3	3	3.00	3.00	0.000
Switzerland	36	1	5	3.00	3.25	1.105
Czech Republic	69	1	5	3.00	3.39	0.771
Total	**116**	**1**	**5**	**3.00**	**3.31**	**0.859**

Alignment of feedback comments with the targeted aspects of the competence

This component focuses on the extent to which students took into account the specified criteria while providing feedback to their peers and whether they drew on those in a thorough manner. Most of the students' feedback comments included all the criteria ($\bar{x}$=3.61, Mdn=4, SD=0.714). Tables 33, 34 and 35 show how the students' feedback comments were scored on this criterion (i.e., number of the criteria in the feedback comments that were addressed) in each round.

Table 33

Number of criteria considered by the assessor in the first round

First Round						
Country	N	Observed minimum	Observed maximum	Median	Mean	Standard Deviation
Cyprus	13	1	3	2.00	2.31	0.630
Switzerland	41	2	4	4.00	3.49	0.675
France (Lyon)	95	1	4	4.00	3.65	0.809
Czech Republic	62	2	4	4.00	3.62	0.657
Denmark	5	2	4	3.00	3.00	0.707
Total	**216**	**1**	**4**	**3.00**	**3.52**	**0.794**

Table 34

Number of criteria considered by the assessor in the second round

Second Round						
Country	N	Observed minimum	Observed maximum	Median	Mean	Standard Deviation
Cyprus	8	2	4	3.00	3.38	0.744
Switzerland	46	2	4	3.00	3.07	0.772
France (Lyon)	96	1	4	4.00	3.69	0.654
Czech Republic	217	1	4	4.00	3.89	0.448
Denmark	51	1	4	4.00	3.47	0.703
Total	**418**	**1**	**4**	**4.00**	**3.69**	**0.637**

Table 35

Number of criteria considered by the assessor in the third round

Third Round						
Country	N	Observed minimum	Observed maximum	Median	Mean	Standard Deviation
Cyprus	11	3	4	3.00	3.45	0.522
Switzerland	38	1	4	4.00	3.76	0.820
France (Lyon)	116	1	4	4.00	3.39	0.832
Czech Republic	69	1	4	4.00	3.61	0.790
Total	**234**	**1**	**4**	**4.00**	**3.52**	**0.814**

An interpretation for this finding is that the use of the various tools, which contained the intended criteria associated with the competence of interest, helped students recognize the aspects of the relevant competence in their peers' artefacts and then to provide feedback based on them. That is, the use of a structured assessment form, based on the given criteria in each case, made it easier for the students to offer feedback comments anchored on the relevant criteria. The provision of assessment criteria increases the probability of establishing a common understanding between students and the teacher of what has to be assessed (Orsmond, Merry, & Reiling, 2000).

In the example below, which relates to the assessment of artefacts associated with the competence of modelling, the assessors had at their disposal an assessment form which included questions about three key features of models, namely representational, interpretative and predictive power (Figure 8).

Model Assessment	
1. Name of model:	Group 9
2. Modeler's group name:	Group 9
3. Date:	24/07/2015
4. Phenomenon represented by your model:	It represents plant growth
5. Criteria 1,2,3,4 : representation of phenomenon **To what extent does the model incorporates all necessary components associated with the operation of the target phenomenon (Criterion 1: objects, Criterion 2: variables, Criterion 3: relationships and Criterion 4: processes)? Specifically identify missing components.**	It includes some objects like plant but some other are missing such glucose and conduits (**CRITERION 1**). Variables like water, sun, CO2 (**CRITERION 2**) and relationships like water-roots, sun – tree leaves, CO2- leaves, are existed (**CRITERION 3**). But it does not provide information about the processes. (**CRITERION 4**)
6. Criterion 5: Explanatory power of the model **To what extent does your peers' model includes a mechanism that can help one account for operation of the target phenomenon? Justify your response and if you think that it does include such a mechanism offer a brief description of the mechanism.**	The model provides a partial explanation. It provides only few pieces of information. (**CRITERION 5**)
7. Criterion 6: Predictive Power **To what extent can your peers' model be used by someone to predict the operation of the phenomenon under specified (not previously observed) conditions? If yes, formulate a prediction and justify how it can be derived from your peers' model**	Even though the plant is already developed, it still needs specific amounts of sun and water. if The plant is irrigated with less amount of water it can't product glucose for its needs. (**CRITERION 6**)
8. How would you suggest your peers to revise/change their model so as to make it more powerful in terms of the six criteria?	They should place the various elements of the model in an appropriate sequence. Specifically, we mean that they have to draw the developments stages of the plan (small, big, bigger) in order to show the size of the plant (**CRITERION 2**) They should also use more colors and be clearer and more explicit in the information they are providing.

Figure 8. Example of student's feedback comments

Another example of a structured template for performing peer assessment, which was used in the trial implementations in Switzerland, is shown in Figure 9. In this example, the students assessed their peers using a variant of a Likert-scale rubric with assessment criteria. The assessors had to choose an appropriate symbol (sad, neutral, happy face) in order to indicate the extent to which the artefact produced by their peers demonstrated acquaintance with a specific aspect of the competence of interest. In addition, the template provided assessors with a space for writing comments.

Another scaffold that was used for facilitating the process of providing feedback comments to peers included an illustrative example of an artefact that could be deemed exemplary (a selected student-generated artefact from previous implementations of the teaching materials or an artefact constructed by the teacher) on the same subject and topic. This given artefact was used by the assessors in order to help them recognize aspects of the competence that were either present in or absent from their peer's artefacts and provide feedback comments, accordingly.

Beurteilungskriterien für "Entladung des Kondensators"

Criteria

Comments

Kriterium	☹	😐	☺	Bemerkungen
Sind die Messwerte so in einer Tabelle dargestellt, dass man erkennt, was gemessen wurde (Welche Werte, Einheiten etc.)?	X			Tabelle fehlt
Ist die graphische Darstellung der Messwerte korrekt? Sind die Achsen korrekt beschriftet? Entspricht die Kurve den Messresultaten (Tabelle)?			X	Übersichtlich mit Farben dargestellt!
Wurden Messungen mit verschiedenen Werten für C und R ausgeführt und graphisch dargestellt?			X	Ja wurde mit vier Möglichkeiten durchgeführt.
Wurden die Halbwertszeiten korrekt aus den Diagrammen herausgelesen und mit den Formelwerten (T = 0.7RC) verglichen?		X		Es wurde nicht rausgelesen, aber gut gerechnet. Noch eine Schätzung schreiben und dann mit den Rechnerergebnissen vergleichen.
Modell: Wird erläutert, welche Elemente des Wasserkreislaufs welchen Elementen des Stromkreises entsprechen (Q, U, I, R, C)?			X	Ja mit Bild ✓
Liefert das Modell eine nachvollziehbare Erklärung für die Beobachtungen am Kondensator (Proportionalitäten zu R und C)?			X	
Wird aus dem Modell klar, was eine Halbwertszeit ist und weshalb die Entladung von Kondensatoren ein exponentiell abfallendes Verhalten zeigt?		X		Noch erklären, was die Halbwertszeit ist, der Rest ist gut.

Figure 9. Example of student's feedback comments

One point that warrants mention relates to one of the implementations held in Cyprus in the first round, where the assessors were not provided with any actual scaffold other than a simple instruction on how to evaluate their peers' artefacts which did not incorporate any specified criteria or structure for creating comments. In this case, the assessors' comments did not capitalize on the relevant criteria. Rather, they emphasized criteria that were neither directly related to the competence of interest nor powerful in terms of helping their peers improve on the relevant competence. Indicatively, these criteria pertained to superficial aspects of the artefact (e.g. use of colours) and grammar or spelling issues. An additional interesting aspect of the data that emerged in this implementation comes from the interviews with the students who explicitly acknowledged the difficulties they encountered in their attempt to provide comments:

> Student: *We prefer to use a form that could guide us because it is more difficult for us to write general comments about the artefacts. Everyone can write anything he wants and sometimes it is not useful for the improvement of the artefact. But if we have specific questions related to criteria, it would be easier for us to give comments to our classmates.*
>
> (Cyprus)

Even though, as mentioned above, in most cases (except for the implementation that did not include any scaffolding) the students' comments considered the pre-specified criteria, they did not exhibit the highest rigor ($\bar{x}$=2.46, Mdn=2, SD=0.903). Tables 36, 37 and 38 show the results from the three rounds.

Table 36

Extent to which the peer-assessor drew on the criteria in a thorough manner in the first round

First Round						
Country	N	Observed minimum	Observed maximum	Median	Mean	Standard Deviation
Cyprus	13	1	3	2.00	2.23	0.599
Switzerland	41	1	4	3.00	2.73	0.895
France (Lyon)	95	1	4	3.00	2.55	0.931
Czech Republic	62	1	4	2.00	2.37	0.854
Denmark	5	1	2	2.00	1.80	0.447
Total	**216**	**1**	**4**	**2.00**	**2.50**	**0.889**

Table 37

Extent to which the peer-assessor drew on the criteria in a thorough manner in the second round

Second Round						
Country	N	Observed minimum	Observed maximum	Median	Mean	Standard Deviation
Cyprus	8	2	4	3.00	2.63	0.744
Switzerland	46	1	4	3.00	2.91	0.755
France (Lyon)	95	1	4	2.00	2.17	0.694
Czech Republic	214	1	4	2.00	2.54	0.912
Denmark	50	1	3	2.00	1.58	0.642
Total	**413**	**1**	**4**	**2.00**	**2.38**	**0.891**

Table 38

Extent to which the peer-assessor draw on the criteria in a thorough manner in the third round

Third Round						
Country	N	Observed minimum	Observed maximum	Median	Mean	Standard Deviation
Cyprus	11	2	3	2.00	2.36	0.505
Switzerland	38	1	4	2.00	2.42	0.948
France (Lyon)	110	1	4	2.00	2.44	0.953
Czech Republic	65	1	4	3.00	2.82	0.983
Total	**224**	**1**	**4**	**2.00**	**2.54**	**0.956**

Reliance of the assessor on the criteria while constructing his/her own artefact

The third component of the coding tool focuses on the extent to which the peer-assessors drew on the intended criteria while constructing their own artefacts. The data suggested that the peer-assessor(s) constructed his/her (their) own artefact by addressing in most cases all the intended criteria ($\bar{x}$=3.29, Mdn=3, SD=0.809). Specifically, Tables 39, 40 and 41 show the number of criteria considered by the assessors in their own artefact in each round in each country.

Table 39

Number of criteria considered by the assessors in their own artefact in the first round

First Round						
Country	N	Observed minimum	Observed maximum	Median	Mean	Standard Deviation
Cyprus	13	3	4	4.00	3.69	0.480
Switzerland	44	3	4	4.00	3.68	0.471
France (Lyon)	98	1	4	3.00	3.02	0.837
Czech Republic	38	2	4	4.00	3.76	0.590
Total	**193**	**1**	**4**	**4.00**	**3.36**	**0. 779**

Table 40

Number of criteria considered by the assessors in their own artefact in the second round

Second Round						
Country	N	Observed minimum	Observed maximum	Median	Mean	Standard Deviation
Cyprus	7	2	4	3.00	3.00	0.577
Switzerland	46	1	4	3.00	2.85	0.729
France (Lyon)	96	1	4	3.00	3.01	0.840
Czech Republic	47	1	4	4.00	3.79	0.587
Total	**196**	**1**	**4**	**3.00**	**3.16**	**0.829**

Table 41

Number of criteria considered by the assessors in their own artefact in the third round

Third Round						
Country	N	Observed minimum	Observed maximum	Median	Mean	Standard Deviation
Cyprus	10	2	4	3.00	3.20	0.632
France (Lyon)	111	1	4	3.00	3.00	0.863
Czech Republic	51	3	4	4.00	3.96	0.196
Total	**172**	**1**	**4**	**3.00**	**3.30**	**0.836**

Additionally, when constructing their own artefact, the majority of the assessors embodied, in a seemingly adequate manner, each criterion in a thorough manner ($\bar{x}$=2.81, Mdn=3, SD=0.815). This implies that students did seem to appreciate how the relevant criteria could be integrated in the artefact that they were asked to develop. Tables 42, 43 and 44 report the outcome of the coding process with respect to the extent to which the peer-assessor(s) draw on the criteria in each round in each country.

Table 42

Extent to which the peer-assessor drew on the criteria in a thorough manner in the first round

First Round						
Country	N	Observed minimum	Observed maximum	Median	Mean	Standard Deviation
Cyprus	13	2	4	3.00	3.08	0.641
Switzerland	44	3	4	4.00	3.68	1.047
Czech Republic	38	2	4	3.00	2.79	0.991
Total	**95**	**2**	**4**	**3.00**	**3.18**	**0.977**

Table 43

Extent to which the peer-assessor drew on the criteria in a thorough manner in the second round

Second Round						
Country	N	Observed minimum	Observed maximum	Median	Mean	Standard Deviation
Cyprus	7	2	4	3.00	3.43	0.535
Switzerland	44	1	4	3.00	2.59	0.726
Czech Republic	46	1	4	3.00	2.91	0.812
Total	**97**	**1**	**4**	**3.00**	**2.98**	**0.768**

Table 44

Extent to which the peer-assessor drew on the criteria in a thorough manner in the third round

Third Round						
Country	N	Observed minimum	Observed maximum	Median	Mean	Standard Deviation
Cyprus	10	2	4	3.00	2.80	0.789
Czech Republic	51	1	4	3.00	2.92	0.659
Total	**61**	**1**	**4**	**3.00**	**2.90**	**0.676**

Provision of sufficient justification to the students about the assessment

The fourth component concerns the justifications provided by the assessor about what the assessee has (or has not) achieved. A feedback comment was considered justified when the assessor provided the assessee with reasons for the credit s/he has given him/her or for the possible weaknesses s/he has alerted him/her for. The justifications of what the students had already achieved for all the three rounds

($\bar{x}$=2.68, Mdn=3, SD=1.392) are approximately equal to the cases of the justifications about what the student had not yet achieved ($\bar{x}$=2.61, Mdn=2, SD=1.436). Tables 45, 46 and 47 are reporting results about the justification provided for the strengths identified in their peers' artefacts, while Tables 48, 49 and 50 report on the justifications about weaknesses.

Table 45

Justification for what students have already achieved in the first round

First Round						
Country	N	Observed minimum	Observed maximum	Median	Mean	Standard Deviation
Cyprus	7	1	5	1.00	1.71	1.496
Switzerland	37	1	5	5.00	4.00	1.374
France (Lyon)	93	1	5	3.00	3.29	1.157
Czech Republic	46	1	5	2.00	2.15	0.894
Total	**183**	**1**	**5**	**3.00**	**3.09**	**1.344**

Table 46

Justification for what students have already achieved in the second round

Second Round						
Country	N	Observed minimum	Observed maximum	Median	Mean	Standard Deviation
Cyprus	6	1	5	2.00	2.67	1.966
Switzerland	38	1	5	4.00	3.66	1.400
France (Lyon)	95	1	5	3.00	3.26	1.044
Czech Republic	204	1	5	1.00	1.70	1.033
Total	**343**	**1**	**5**	**2.00**	**2.37**	**1.370**

Table 47

Justification for what students have already achieved in the third round

Third Round						
Country	N	Observed minimum	Observed maximum	Median	Mean	Standard Deviation
Cyprus	9	1	5	4.00	3.22	1.922
Switzerland	30	1	5	2.00	2.60	1.694
France (Lyon)	106	1	5	3.00	3.10	1.218
Czech Republic	43	1	5	2.00	2.19	1.160
Total	**188**	**1**	**5**	**3.00**	**2.82**	**1.376**

Table 48

Justification for what students have yet to achieve in the first round

First Round						
Country	N	Observed minimum	Observed maximum	Median	Mean	Standard Deviation
Cyprus	13	1	5	3.00	2.69	1.601
Switzerland	39	2	5	5.00	4.79	0.656
France (Lyon)	89	1	5	3.00	2.64	1.299
Czech Republic	35	1	5	2.00	2.34	1.162
Total	**176**	**1**	**5**	**3.00**	**3.06**	**1.501**

Table 49

Justification for what students have yet to achieve in the second round

Second Round						
Country	N	Observed minimum	Observed maximum	Median	Mean	Standard Deviation
Cyprus	7	1	5	2.00	2.29	1.496
Switzerland	40	1	5	5.00	4.33	1.118
France (Lyon)	92	1	5	2.00	2.37	1.146
Czech Republic	189	1	5	1.00	1.74	0.936
Total	**328**	**1**	**5**	**2.00**	**2.24**	**1.320**

Table 50

Justification for what students have yet to achieve in the third round

Third Round						
Country	N	Observed minimum	Observed maximum	Median	Mean	Standard Deviation
Cyprus	11	1	5	5.00	4.00	1.612
Switzerland	30	1	5	4.00	3.30	1.803
France (Lyon)	106	1	5	3.00	2.74	1.290
Czech Republic	41	1	5	2.00	2.27	1.141
Total	**188**	**1**	**5**	**3.00**	**2.80**	**1.430**

The comments below are examples of references about what the students have already achieved (first example) or not achieved (second example), which did not incorporate a sufficient justification that could be informative for the students.

> *According to their calculation it is correct.*
>
> *But the triangles in the row were incorrectly counted; you should count only the bottom triangles.*
>
> (Czech Republic)

Provision of guidance

The final component of the reporting template focuses on the guidance provided by the assessor(s) to the assessee(s). It relates to (i) the extent to which the feedback was specific enough, suggesting concrete next steps that could be taken for improving their artefact and (ii) the extent to which the guidance is indeed related to the competence of interest. Specifically, the guidance provided by the majority of the assessors was rather vague ($\bar{x}$=2.33, Mdn=2, SD=0.972), often lacking clear suggestions for concrete next steps. Additionally, the assessors' guidance was mostly related to the competence of interest and not to other aspects such as the appearance of the artefact ($\bar{x}$=0.89, Mdn=1). Tables 51, 52 and 53 summarize the outcome of the coding of this aspect of the student-generated feedback. An example of a rather specific guidance provided by an assessor is the following:

> *We do not understand the procedure, because it will not be possible when the shapes are more complicated. They had to make marks or connect those parts that belong to each other.*
>
> (Chech Republic)

Table 51

Guidance provided by assessors in the first round

First Round						
Country	N	Observed minimum	Observed maximum	Median	Mean	Standard Deviation
Cyprus	13	1	4	2.00	2.31	0.855
Switzerland	40	1	4	2.00	2.45	0.783
France (Lyon)	93	1	4	2.00	1.87	0.811
Czech Republic	42	1	4	2.00	1.95	0.909
Total	**188**	**1**	**4**	**2.00**	**2.04**	**0.858**

Table 52

Guidance provided by assessors in the second round

Second Round						
Country	N	Observed minimum	Observed maximum	Median	Mean	Standard Deviation
Cyprus	8	1	4	2.00	2.00	0.640
Switzerland	31	1	4	2.00	1.87	0.856
France (Lyon)	95	1	4	2.00	1.83	0.741
Czech Republic	201	1	4	1.00	1.15	0.443
Denmark	36	1	3	2.00	1.83	0.811
Total	**371**	**1**	**4**	**2.00**	**2.54**	**0.876**

Table 53

Guidance provided by assessors in the third round

Third Round						
Country	N	Observed minimum	Observed maximum	Median	Mean	Standard Deviation
Cyprus	11	2	3	3.00	2.82	0.982
Switzerland	30	1	4	4.00	3.70	0.794
France (Lyon)	110	1	4	2.00	1.87	0.968
Czech Republic	58	1	4	1.00	1.97	1.123
Total	**209**	**1**	**4**	**2.00**	**2.21**	**1.178**

3.4.2 Assessors' actions

The available data seem to provide insights into certain aspects, which relate to the productiveness of the students' implementation of the two key roles (i.e., assessor and assesse). In most cases, students' effort to provide feedback was supported by a specially designed tool, which explicated the relevant criteria for the competence of interest. Thus, the assessors were supported to systematically judge their peer's artefact against each criterion.

An indication that could provide insights as to whether the students enacted productively the role of the peer assessor refers to the extent of internal consistency. In particular, while developing their own artefact, they were asked to address certain criteria, associated with the competence of interest. These criteria were negotiated beforehand with the teacher and were coded as criteria to be satisfied while producing the artefacts (as part of the teaching/learning materials and the respective activity sequence). One interesting indication that we focused on, relates to the extent to which these criteria were also employed by the students while assessing the artefacts produced by their peers. For this, we sought to identify whether the students who had addressed those criteria while developing their own artefact,

also drew on these same criteria for evaluating the artefacts created by their peers. The data show that the students who had attended to certain criteria while developing their own artefact, did not necessarily attend to those criteria while assessing their peer's artefact. One possible interpretation for this apparent inconsistency is that students did not really appreciate the need to commit to certain criteria while assessing their peer's artefacts. An alternative interpretation is that they did not appreciate that coordinating the process of the development of their own artefact with the process of assessing their peer's artefacts, could serve to systematise and facilitate the latter. This provides an indication of a potentially non-productive stance towards the role of peer-assessor and highlights the need to help students frame the process of peer-feedback in an appropriate manner that would render it meaningful.

In addition to the cases where the students avoided employing certain criteria while assessing their peer's artefacts, even though they seemed to appreciate the substance of that criterion, we also encountered cases that run in the reverse direction: students who avoided addressing certain criteria while developing their own artefact employed those same criteria while assessing their peer's artefacts. This is admittedly not surprising. It might just be that students did understand the importance and the essence of a certain criterion, but they were not well positioned to actually apply it in the given context. Put differently, this could be signalling limited conceptual understanding about the content associated with the given situation rather than lack of understanding about the importance of criteria in enacting the role of the assessor. However, one noteworthy aspect of this finding is that the implementation of the role of the peer assessor could serve as a valuable teaching and learning experience for the students (as peer assessors) since they could identify possible ways to improve their own artefact. This is illustrated in the following excerpt from an interview with a student[4]:

> *Interviewer: In your own artefact you did not mention any predictions of how the plants are growing. However, you have provided suggestions to your peers about how their artefact could be improved. Were you able to do this, since this criterion was not included in your model?*
>
> *Student: Yes, indeed. But when we observed our peers' artefact (model), we understood how we could add predictions to our own model too. We had an example of how we could do this.*

3.4.3 Assessees' actions

Another important aspect of the peer feedback process relates to the assessees' reaction to the feedback comments they received from their peers. We tried to probe this through the interview data from all three rounds. This allowed us to discern two certain ways of dealing with the feedback comments.

[4] The interview has been reported by a LWG in Cyprus and the excerpt concerns the criterion of interpretation of the modelling competence. Specifically, the criterion focuses on mechanism of operation of the target phenomenon.

The first way involves using the assessors' feedback comments to improve the artefact that was produced. Indeed, there were instances where the assessees saw merit in the feedback comments they received, identified those with specific weaknesses characterizing their artefact and undertook corresponding refinements to alleviate those shortcomings. The following example is related to the competence of modelling and specifically the interpretation of how plants grow (criterion 5):

> *Interviewer: Did you find the feedback comments from the assessors' group about criterion 5 useful?*
>
> *Student: Yes. The comments were very useful. They helped us to identify weaknesses of our model and also what we could do in order to improve the interpretive ability of our model.*

The second way includes avoiding to use seemingly productive feedback comments i.e. feedback provided about what the assessees had yet to achieve with sufficiently elaborated justifications, accompanied by guidance about concrete next steps on how to go about it. The interview data provided indications for possible reasons underlying this rather unproductive stance. One of these reasons is that the assessees could have perceived their peers as non-legitimate assessors. For instance, if assessees thought that their assessors were not well positioned in terms of the competence of interest, they were likely to dismiss their feedback comments from the outset. The following examples illustrate this. In the two first cases, the assessees received feedback about the criterion of prediction on the competence of modelling (this was the sixth of the criteria for what constitutes a powerful model)[5]:

> *Interviewer: We noticed that you did not use the guidance for criterion 6 provided by assessors. Why did you select to not use it?*
>
> *Student: Because their artefact was not good, so their feedback was not going to be any good either. If their artefact is missing a lot of things, how could they possibly give us useful feedback? Our model was better anyway.*
>
> *Student: The assessor's model was not good so they could not provide us with reliable and useful feedback.*
>
> *I prefer feedback from a teacher because I can trust her", "The teacher knows more things and she is smarter than my peers"*[6]

Furthermore, in the case of an implementation that took place in the Czech Republic, 47% of the 34 participating students, preferred teachers' comments as compared to the 29% of the students who mentioned that they found their peers' comments more useful. Specifically, they stated that "*the teacher is more clever and teaches us everything" or "when my peer assesses me, I do not give particular emphasis on his comments because he is a friend but the teacher is the person who will give me a*

[5] The interview is from a Cypriot LWG and the excerpt concerns the criterion of prediction, a part of the modelling competence. Specifically, the criterion focuses on certain conditions which may influence the operation of a phenomenon in a future moment.

[6] The interview is from a Czech Republic LWG

grade". From these statements we can conclude that students commonly trust the teacher's feedback comments because (s)he has much more concrete background knowledge, as compared to their peers. In addition, they were influenced by the fact that the teacher has the authority to give marks, which makes the teacher's comments more significant.

A further interpretation for this stance might be related to the social aspect of the interaction, in terms of the relationships between the peers. It might be that this social aspect could incur an emotional load to the students, which tends to undermine the process of genuinely assessing their peer's artefact (or feedback comments) in a critical but also objective manner. The excerpts below pertain to a conversation between the researcher and a student who did not act on their peers' comments about the processes that become relevant to plants' growth (criterion 4: specification of relevant *processes*)[7] and the mechanism underlying these processes (criterion 5: incorporation of a *mechanism underlying the operation of the model*)[8] respectively:

> Interviewer*: We noticed that you did not use the guidance you received from your peers for criterion 4. Why didn't you use it?*
>
> Student A: *The students who assessed our model did not seem to like us. So, we did not expect them to write useful feedback. At the end, they asked us to change several things in our model that we had already changed. So their comments were not reliable.*
>
> Student B: *Seriously, when I read our peer's feedback comments I was shocked. I think they didn't like us because they asked us to change many things. Our model did not need so many improvements. I repeat, they don't like us.*

This interpretation also surfaced in the teacher's self- reported data, as shown in the following excerpt:

> *Students are afraid of biased assessment from their peers (mainly based on the relationships between them and not based on the quality of the work). It is not possible to provide an environment in the classroom where students can't recognize who is assessing them.*
>
> (Czech Republic)

These findings suggest a specific challenge that needs to be addressed in any attempt to enact peer-feedback as an assessment method. In particular, peer-feedback posits that students have developed certain skills, which are needed to enable effective implementation of the relevant roles (assessors and assessees). Also, it posits students' appreciation of peer-feedback as a useful, productive exercise that could facilitate learning. This is needed to help them adopt a useful and productive stance towards this process. In any case, fulfilling these requirements posits acquainting students with multiple

[7] The interview is from a Cypriot LWG and the excerpt relates to the aspect of representation: associated criteria include the specification of the objects, variables, relationships and processes which become relevant to the phenomenon to be modeled. Specifically, this interview focuses on the fourth criterion (processes).

[8] The interview is from a Cypriot LWG and the excerpt relates to the aspect of interpretation which is associated with the incorporation of a mechanism underlying the operation of the model. The specification of an appropriate mechanism is one of the criteria that were used for assessing the quality of models in this intervention.

opportunities to engage in this process and enact the two roles. Also, this could be usefully supplemented with explicit discussion as to the features of what could be deemed useful feedback. Of course the importance of acquiring experiences with what constitutes effective feedback is not restricted to the students; it is also of paramount importance for teachers themselves.

4. Assessment method: On the fly Interactions

4.1. Information on how on the fly Interactions was incorporated in the teaching intervention

The assessment method "Interactions On the fly" was tested in KCL (United Kingdom), UGA (France, Grenoble), JYU (Finland) and UCY (Cyprus). In total, in all three rounds, this assessment method was implemented in 40 cases. Tables 54, 55 and 56 summarize the parameters associated with each of these 40 implementations in terms of the targeted competence, educational level, number of participating students, and subject and specific topic in which the teaching materials were situated, for the corresponding round of implementation.

With respect to competence, the majority of the implementations were associated with investigation (30) while the remaining ten were associated with problem solving. Regarding educational level, sixteen of 40 implementations were implemented in lower secondary, fifteen in upper secondary grades and the remaining nine in primary education. The implementations covered different domains of Natural Sciences such as Physics (15), Biology (5), Integrated Sciences (2) and Mathematics (13) and Technology (5) (Figure 10).

Table 54

Implementations based on "On the fly Interactions" in the first round

	Competence	Educational level	Number of students	Subject	Topic	Country
1	Investigation	Lower Secondary	20	Mathematics	Fractions writing and Calculus	France (Grenoble)
2	Investigation	Lower Secondary	16	Physics	State of matter change/ Water Temperature evolution	France (Grenoble)
3	Investigation	Lower Secondary	16	Technology	Constructing a bridge with papers	France (Grenoble)
4	Investigation	Lower Secondary	17	Physics	Electricity	Finland
5	Problem solving	Primary	23	Mathematics	Arithmetics	Finland
6	Problem solving	Upper Secondary	26	Mathematics	Geometry	Finland
7	Investigation	Upper secondary	12	Physics	Kinematic-Free Fall, Spring elongation	Cyprus
8	Investigation	Upper Secondary	18	Physics	Forces	United Kingdom
9	Investigation	Upper Secondary	27	Physics	Electromagnets	United Kingdom
10	Problem solving	Lower Secondary	15	Mathematics	Triangle numbers	United Kingdom
11	Problem solving	Lower Secondary	30	Mathematics	Fractions Arithmetic	United Kingdom
12	Investigation	Lower Secondary	30	Biology	Micro-organisms	United Kingdom

13	Investigation	Primary	30	Biology	Micro-organisms	United Kingdom

Table 55

Implementations based on "On the fly Interactions" in the second round

	Competence	Educational level	Number of students	Subject	Topic	Country
1	Investigation	Lower Secondary	20	Mathematics	Fractions writing and Calculus	France (Grenoble)
2	Investigation	Lower Secondary	16	Physics	State of matter change	France (Grenoble)
3	Investigation	Lower Secondary	16	Technology	Materials properties / Constructing a paper bridge	France (Grenoble)
4	Investigation	Primary	33	Science	Forces	United Kingdom
5	Investigation	Upper secondary	27	Physics	Electromagnetism	United Kingdom
6	Investigation	Primary	30	Science	Making bubbles	United Kingdom
7	Investigation	Primary	30	Biology	Micro-organisms	United Kingdom
8	Problem solving	Primary	11	Mathematics		United Kingdom
9	Problem Solving	Lower secondary	32	Mathematics	Triangle numbers	United Kingdom
10	Problem Solving	Primary	50	Mathematics	Geometry	Finland
11	Investigation	Lower secondary	45	Physics	Waves and thermodynamics	Finland
12	Problem Solving	Upper secondary	65	Mathematics	Statistics & Integration	Finland
13	Investigation	Upper Secondary	16	Physics	Kinematics-Free Fall, Spring elongation	Cyprus
14	Investigation	Upper	17	Physics	Kinematics-Free Fall,	Cyprus

		Secondary			Spring elongation	

Table 56

Implementations based on "On the fly Interactions" in the third round

	Competence	Educational level	Number of students	Subject	Topic	Country
1	Investigation	Upper secondary	21	Physics	Kinematics-Free Fall, Newton's laws	Cyprus
2	Investigation	Upper secondary	20	Physics	Kinematics-Free Fall, Newton's laws	Cyprus
3	Investigation	Upper secondary	30	Physics	Floating orange	United Kingdom
4	Investigation	Upper secondary	14	Biology	Membrane permeability	United Kingdom
5	Investigation	Upper secondary	30	Physics	Floating orange inquiry-SAILS	United Kingdom
6	Investigation	Upper secondary	14	Biology	Membrane permeability	United Kingdom
7	Poblem solving	Lower secondary	16	Mathematics	Volume cuboids	United Kingdom
8	Problem solving	Upper secondary	28	Mathematics	Pair Products	United Kingdom
9	Investigation	Lower secondary	30	Science	Separating	United Kingdom
10	Investigation	Primary	30	D&T in science	Healthy Crisps	United Kingdom
11	Investigation	Lower secondary	16	Technology	Constructing a bridge with peers	France (Grenoble)
12	Investigation	Lower secondary	20	Mathematics	Calculus/fractions	France (Grenoble)
13	Investigation	Lower	16	Physics	State of matter	France

	secondary	changes / Water Temperature evolution	(Grenoble)

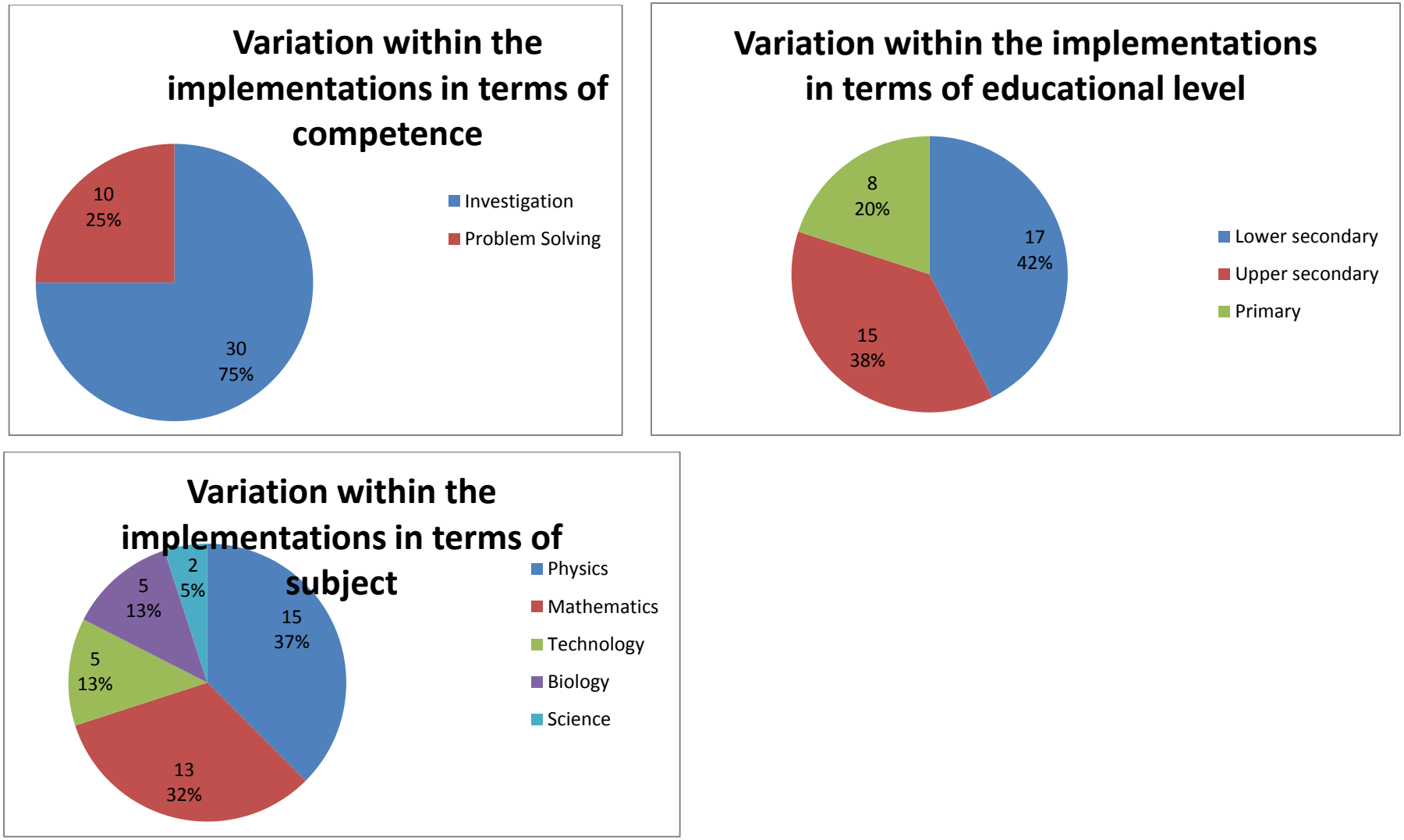


Figure 10. Parameters associated with the implementations of all the three rounds

4.2 Research Data

In line with the research design that refers to this assessment method, the LWGs undertook to develop learning environments that would ensure ample opportunities for either whole class (in all cases in Cyprus and in some cases in Finland) or group discussions (in all cases in the UK and in some cases in Finland). To some extent, these discussions were planned, in the sense that the teacher identified beforehand issues that s/he should be aiming to bring into focus. However, their actual content or structure was not planned in any way and it was up to the teacher to steer the discussion by managing and coordinating the spontaneous interactions that unfolded in the class.

In all cases, the lessons were videotaped, or audiotaped, and those parts of the video that included interactions "on the fly" were identified and transcribed. Field notes were taken during the inquiry lessons. An additional data source involved interviews with the teachers, after the implementations. The interview sessions, were often stimulated by means of video recorded episodes of the corresponding classroom discussions. Specifically, the teachers watched pre-selected parts of the video recordings of their lessons, as a means to elicit useful reflection, on the part of the teacher, about particular issues, associated with aspects of the specific assessment method that emerged during the discussion. Moreover, the teachers filled in a self-reporting tool in relation to the assessment method they had implemented (Appendix III).

Figure 11 illustrates the research design. The actual research design (Appendix IV) provides further information about the a) focus of the research design, b) specific research questions associated with the assessment method, c) rationale, and d) constraints to be satisfied.

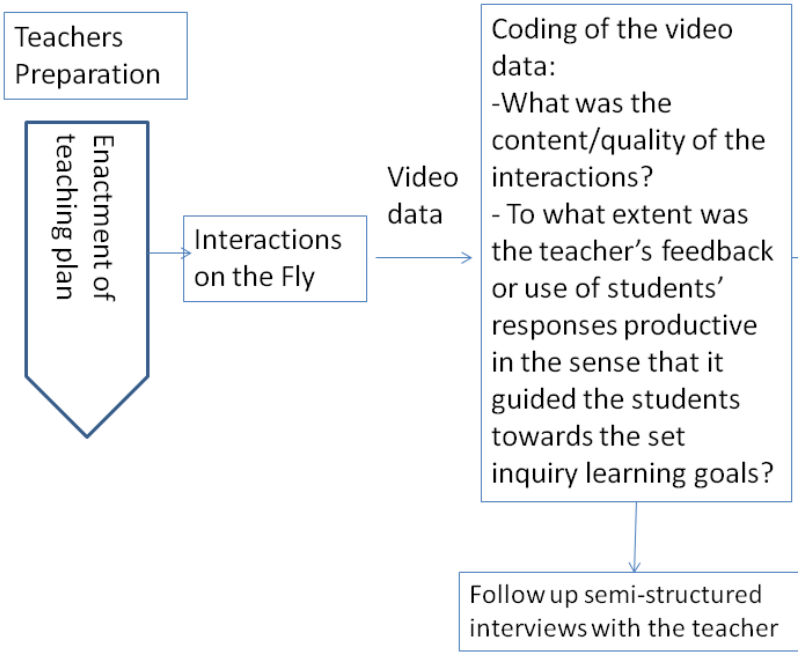


Figure 11. Illustration of Research Design I

4.3 Data Analysis

The data arose from teacher-student interactions either during the course of an inquiry activity or from class or group discussions within inquiry lessons. The data that emerged from the classroom implementations were processed so as to investigate what seems to facilitate or impede the effectiveness of this assessment method. Following on from this, we then explored the challenges that seemed to come into play when cycles were broken, as well as possible ways to address them.

For this purpose, interactions between teachers and students were coded using the ESRU scheme (Ruiz-Primo & Furtak, 2006). Specifically, the interactions were coded as either instances of E (elicit), S (students' response), R (recognize) or U (use)[9]. The following examples illustrate how the various contributions in dialogue were coded.

In this example, the teacher tries to introduce the concept of force as an instance of interaction between objects in a class of 15-year-old students. In this part of the dialogue the teacher seeks to introduce and negotiate the meaning of a central idea of the teaching unit, namely the "environment of an object", which is intended to refer to the set of objects that interact with the object of interest by means of forces. The teacher seeks to promote discussion intended to facilitate consensus on the meaning of this term.

T: What else could be considered as part of the environment for this object?	*E*
S: The ground.	*S*
T: The ground...	*R*
S: The Earth	*S*
T: What is the difference between ground and Earth?	*(R)U*[10]

Ruiz-Primo and Furtak's approach suggests that complete ESRU cycles and iterations of complete ESRU cycles could provide an indication of potentially productive interactions between the teacher and the students in the assessment dialogue framework (Ruiz-Primo & Furtak, 2006). Our aim was to explore the cycles identified in our data sets, to focus on what seemed productive dialogue that promoted the inquiry process and to identify instances in which the ESRU cycle happened to break. We also wanted to elaborate on possible reasons underlying each of these different instances. Our aim was to identify and describe the variation within the ESRU cycles. We were also interested in what caused ESRU cycles to break and identify whether this resulted from pedagogic decisions by the teacher or highlighted areas where there were missed opportunities in feedback that impeded taking

[9] An optional component of the research design involved further coding these utterances into subcategories so as to identify and describe the variation within these main categories, e.g., the different ways of eliciting information by the students or different ways of using contributions made by the students to sustain, deepen or focus the discussion. The outcome of this optional analysis appears in Appendix III.

[10] This utterance was double coded since it reflects an instance of both recognizing and using inputs from a student.

learning forward. This can provide evidence of challenges and intricacies associated with interactions on the fly as a method of formative assessment.

The UK and Finnish partners found that, in some cases, the ESRU cycle analysis was not sufficient in capturing all aspects of the quality of some of the inquiry dialogue and so developed an alternative coding system formulated from work on convergent and divergent formative assessment approaches (Torrance and Pryor 2001). This further analysis enabled the UK and Finnish teams to focus on incidents in the classroom dialogue where the teachers either utilised questioning to open up student thinking to further ideas (divergent assessment) or to close down the thinking to either reach a consensus or provide more direction for the student (convergent assessment).

Table 57

Convergent/Divergent codes adapted from Torrance & Pryor (2001)

Codes	Description
TD – Divergent	Questions/comments that open up discussion (how/why?); probing thinking
TC – Convergent	Questions/comments to check if students are on track; focusing the discussion
TL – Lectures	Give information through exposition or direct instruction
TA – Affirms	Acknowledge/repeat/re-formulate student(s) contribution
TN – Inquiry	Make explicit reference to inquiry (general principles or procedures) and to NoS

The following excerpt provides an example of application of KCL coding system. It is the same excerpt as presented above.

Turn	Speaker	Transcription	Code
1	Jane	So what you get?	TD
2	Student	It is all kind of awkward…	
3	Jane	Go on	TD
4	Student	So I don't know, it is as if …	
5	Jane	So that could be a straight line or it could be a relationship that sort of tails to that end.	TC
		How would you know?	TD
		What could you do to improve it?	TD
		Where are you missing data?	TC
6	Student	Here (points at results)	
7	Jane	Yes, fill those gaps in and those gaps in and then you would know.	TL

Figure 12 illustrates the pattern of the conversation presented above combining ESRU and KCL coding systems. ESRU codes are in the plot area in temporal sequence and KCL codes in y-axis.

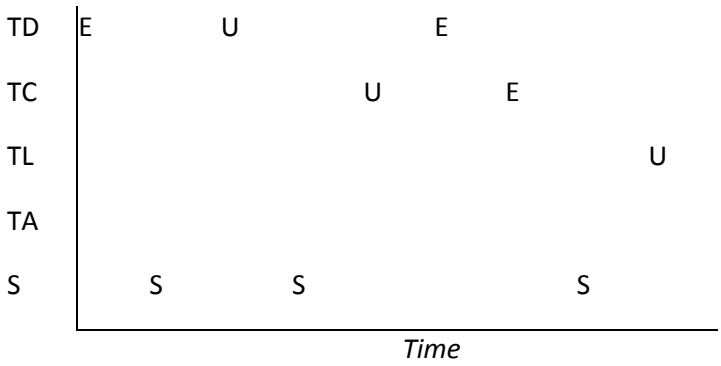


Figure 12. Pattern emerging from the analysis of one interaction on-the-fly taken from a physics inquiry lesson in England.

Data collection and coding were carried out by the researchers in the corresponding LWGs. Toward this end, the LWGs were provided with relevant templates for coding and reporting both the data from the video and audio recordings but also the interview data (Appendix III). In the next section, we present the results that emerged from the synthesis of the data that were coded and reported by the individual LWGs.

4.4 Results

The section is structured in three parts. These parts correspond to the research questions specified in the relevant research design. The first deals with the patterns that can be identified in the interactions between the teacher and the student. The second focuses on the researchers' interpretations of the teachers' formative actions in response to the feedback they were receiving from their students. The third summarizes the findings from the interview data, concerning teachers' perceptions about the challenges and opportunities associated with "*interactions on the fly*" as an assessment method.

Patterns that can be identified in the teacher-students interactions

The coding of 11 instances of classroom dialogues during the three rounds of interventions in Cyprus, revealed a total of 189 completed ESRU cycles (51% of the instances in which such a cycle started) and 182 broken ESRU cycles (Table 58). Coding the 9 instances from the 3 rounds of implementations in France (Grenoble) revealed a total of 200 completed ESRU cycles (40% of the instances in which such a cycle started) and 294 incomplete ESRU cycles. In a similar manner, coding 18 instances of dialogue from the teaching interventions in Finland, revealed that the ESRU cycle was completed 135 times, a number that corresponds to 56% of the instances in which such a cycle started. In the remaining 44% of the cases (185

instances) the cycle remained incomplete, either as ESR or ES. It is worth noticing that the percentages of complete and incomplete ESRU cycles in these countries are rather similar.

Table 58

Complete and incomplete ESRU cycles.

ESRU cycles	UCY	Grenoble	Finland
Complete	189 (51%)	200 (40%)	235 (56%)
Incomplete	182 (49%)	294 (60%)	185 (44%)

Factors affecting the ESRU Cycles

The available data provided indications about possible reasons why certain instances of ESRU cycles were successfully completed. We were able to identify six key features, which seemed to have led to the completion of ESRU cycles. These features are summarized in Table 59.

Table 59

Variation of complete ESRU cycles.

	Complete ESRU cycles
1	Teacher suggests an activity in order to investigate a hypothesis that emerged during the discussion.
2	Teacher poses a question that is intended to promote students' thinking about the topic being discussed
3	Teacher poses clarification questions to help students further articulate a seemingly valid contribution they made
4	Teacher takes the opportunity to use what the students had said in order to conclude/provide feedback
5	Students expressed several ideas that allowed the teacher to ask for comparisons between these ideas
6	Teacher gets a contribution from the students and asks for further explanation with the intent to help them evaluate and adjust their reasoning

In each of these six cases, it is evident that the teacher is seeking to promote further thinking by the student or students, which is likely to lead to richer dialogue. Sometimes this allowed the teacher to obtain more feedback from students about their reasoning. In other cases, it served to challenge students' understanding so that they could adjust the direction in which the inquiry was developing. In yet other instances, the teacher raised questions to deepen and continue the dialogue between students and student groups so that the ideas emerging from these scenarios could serve as both guidance and as a resource for other students.

Similarly, we categorized our interpretations for the cases where information that came up in the discussion did not seem to be used in an explicit manner and led to incomplete ESRU cycles (Table 60).

Table 60

Variation of incomplete ESRU cycle

	Incomplete ESRU cycles
1	Teacher gets the answer s/he had been looking for but s/he doesn't ask for further explanation, while it would have been useful to do so
2	Teacher ignores/rejects/interrupts student's contributions to the dialogue (possible because s/he either considered them irrelevant/wrong or was looking for a different contribution)
3	Teacher offers the right answer or guides students to say the correct word instead of promoting their thinking or questioning their ideas
4	Teacher poses a vague question that students might have not been able to construe in a meaningful manner
5	Teacher poses a closed yes/no question
6	Teacher rejects an answer as non-compatible with observation/accepted theory that s/he wants them to focus on
7	Teacher does not seem to pay adequate attention to a student who tries to express a relevant/potentially interesting question, having the intention to move to another subject

The cases above highlight instances when the dialogue either halted or did not reach fruition resulting in broken ESRU cycles. There were also instances where the cycle was broken but the discussion got redirected or ideas got picked up later. In these latter broken cycles, it was clear that the teacher, having provided thinking time both for themselves and for further deliberation by the students, redirected the talk back to pick up unfinished ideas.

So while sometimes the breaks in the cycle indicated instances where teachers manufactured more thinking time within the flow of inquiry ideas, there were also instances where the breaks in cycles identified points where there might have been more opportunity to guide or consider ways forward.

The notion of challenges

In an attempt to shed more light into the intricacies underlying the teachers' attempts to employ interactions on the fly as a formative assessment method, we focused on instances where either important information (i.e., contributions made by the students) went unnoticed during the discussion or was not used in a seemingly optimal manner. As part of the coding process, the LWGs were expected to seek and report such instances. We then undertook to synthesize the findings reported by individual LWGs with the intent to come up

with a typology of such challenges. Next, we describe the different types of challenges that were reported, at the collective level.

Challenge 1

The teacher is exposed to various inputs made by the students (usually as a response to his/her own probes). Some of these contributions are more likely to support the evolution of the classroom dialogue in a productive manner. At the same time, it is always possible that elaborating on (even apparently useful) contributions might end up taking the discussion away from the inquiry focus for that activity. These two features pose an important challenge for teachers, which entails two aspects: (a) how to make optimum decisions in real time as to which students' contributions to draw upon, in the sense that they are more likely to lead to productive discourse and what aspects to suppress or ignore (at least temporarily) (b) ensure a stance that is responsive/attentive to students' contributions but also allows steering the discussion in an effective manner, toward the learning goals (e.g. offer explicit guidance/feedback, deepen the discourse around the selected students' ideas through reflective questions, narrow down the scope of the discussion etc). Making decisions, in real time, about his/her immediately next moves in the evolving dialogue, constitutes a very challenging and demanding task for the teacher.

Next we present some examples that serve to better illustrate the first aspect of this particular challenge, which relates to the decisions the teacher needs to make of which students' contributions to elaborate more and which to suppress.

Example 1: In some cases, the teacher encounters the dilemma whether to address certain issues that come up during the discussion and could have a profound influence on the discussion, or to strictly confine himself/herself within the lesson plan. This is illustrated in the following excerpt from the classroom dialogue about the forces that are applied on an object placed on a table:

> *T: What can we think (which forces are applied on the object)?*
> *S1: The earth and the table apply forces on the object.*
> *S2: And the object applies a force on the table.*
> *T: Let's leave that for now. What did we say we are interested in at this time?*
> *S1: The forces that are applied on the object.*

In this case, the teacher seems to choose not to extend the discussion to include the idea of action-reaction. This idea (action/reaction) seems to be underlying how these students actually conceptualize the subject under discussion. In particular, even though there is a clear distinction for the teacher between the "forces being exerted by an object" and "the forces being experienced by an object", this distinction might not be clear for the students and is therefore likely to underlie the discussion the whole time. This is evidenced by the fact that this specific idea emerged repeatedly at different points in the classroom discussion. For instance, the excerpt below shows how this idea surfaced again later in the

same episode (i.e., while discussing about the forces applied on an object and the direction of each).

> *T: So this force is due to the direct contact with the table. Because the table is in contact with the object, some said that it exerts a force downwards and others upwards. Who is correct?*
> *S: Since the object is still, it means that the body exerts a force on the table and the table exerts a force of the same magnitude on the object.*
> *T: You are right, but these forces are not exerted on the object. We are only interested in the forces exerted on the object. Which are these?*

The selection to avoid discussing this particular idea (action-reaction) could reflect a strategic choice made by the teacher. One possible reason for this strategic choice could be related to the fact that action-reaction was to be addressed in a subsequent lesson, and the teacher might have selected not to deviate from the current lesson plan prescribed by the textbook. This interpretation seems to be consistent with the interview data with this same teacher. In particular, this teacher was asked to reflect on her choice to not address the notion of action-reaction and she stated that *"we will do such examples in a later lesson when we will discuss Newton's third law".* In any case, making such decisions is a challenge for teachers.

<u>Example 2</u>: The next excerpt shows the classroom discussion about the results found in an experiment undertaken by the students in which they measured the acceleration of an object, when released from different heights. One challenge associated with this example relates to the dilemma as to whether to deal with students' difficulties with certain aspects of experimentation (systematic/random error and possible sources of error in a given experiment) or, to keep the focus on the conceptual aspect, namely the consistency between the experimental results and the corresponding theory.

> *T: What is your conclusion?*
> *S1: When the height increases, the acceleration drops.*
> *T: By how much does it drop? This 8.4 could be 8.5 so if we just ignore the decimal part of the number this would round up to 9. This is also the case for these two measurements (points to other 2 measurements).*
> *S2: So they are the same…*

It is important to note that the notion of experimental error and its distinction from systematic variation was one of the goals of the lesson, as these were described by the teacher. Yet, in this instance, the teacher avoided discussing experimental errors, how they might emerge and how they could influence measurements. Rather, she skipped this potentially useful discussion to illustrate for students how the observed variation in the measurements of acceleration is sufficiently small to be dismissed as insignificant. In particular, she suggested that putting aside the decimal parts of the numbers, the measurements seem to round up to the same whole number, which implies that height does not influence acceleration.

Another example from this same lesson relates to the discussion about the results from a set of experimental trials where objects of different mass were released from the same height. At some point in this discussion a student had noticed that they were essentially extending the scope of the validity of a presumed causal relation they had detected between mass and acceleration, to a range of values other than those they had actually explored. He then expressed his reservations about the validity of this extrapolation from the data.

> *S1: Mass has an influence because they put 7gr and we put 13gr.*
> *S2: But mass doesn't have an influence.*
> *S1: It does influence.*
> *T: So it does… But here we found the same acceleration whether we put 4gr, 7gr or 14gr.*
> *S1: We added only 4gr, how would that make any difference?*
> *T: What are you saying? Does mass influence?*
> *S2: If we had added 1 Kg it would still be the same, yes?*
> *T: If we had added 1 Kg would it still be the same? Why does it have to be the same?*
> *S2: Because the acceleration of gravity is the same.*
> *T: That is what we wanted to prove. We know it, but we wanted to prove it.*

The student's objection is reasonable; certain pairs of variables might indeed have a different relation at different ranges of values (e.g. the force that extends a spring with the extension of the spring). However, the teacher focuses on what is correct according to canonical physics knowledge, without discussing this issue. The challenge here for the teacher is to balance priorities between focusing to conceptual meanings or devoting time for addressing aspects of scientific processes or epistemological issues that emerge in discussion.

<u>Example 3</u>: Students work in groups to develop hypotheses about the behavior of oranges in water, and verify their hypotheses by experimentation. The following excerpt of classroom dialogue relates to variables that might influence sinking or floating of an orange in water.

> *T So now, go back to your question, when I squeeze the pith, what may be…*
> *S* ***Is that up thing?***
> *T So, what must be changing, or were you changing when you squeezed that one that used to float and then it sinks, what must you be changing, by deduction of what we just talked about, you are changing its…?*
> *S Mass.*
> *T No, sorry, no, no, I don't know would be the right answer, what scientific context are you looking at when you look at floating and sinking, tell me, you've already said it to me, it's a D word.*

The student above made a vague reference to a potentially important and highly relevant idea ("up thing"), which the teacher avoided to explicitly probe further. In this context, it might have been more effective for the teacher to ask follow-up questions (e.g. what do you

mean by…?) so as to initiate useful discussion that would allow articulating important ideas. Admittedly, it is not possible to discern whether the teacher noticed this opportunity but thought that it would have been better from an instructional perspective to avoid explicitly addressing Archimedes principle at this stage (e.g. so as to first elicit more ideas from the range of the class) or whether s/he did not actually notice it. In any case, this example could serve to illustrate a situation where the teacher confronts with some load of information that is not easy to manage and the need to make choices as to which students' contributions to draw upon, in the sense that they are more likely to lead to productive discourse and what aspects to supress might not be the optimal ones.

Example 4: In this episode students (S2 & S4) ask two questions which are not answered by the teacher. Especially, the question expressed by S4 could be fruitful for classroom discussion, but the teacher avoided to bring them into the spotlight and instead selected to move forward to another topic.

> *T: What do those blue balls represent over there? [Refers to a simulation about electric current.] Student1?*
> *S1: Electrons*
> *T: Yeah. Maybe this causes a mental image that there are really some balls travelling. Travelling like…*
> *S2: Well, when do they run out?*
> *T: It runs out when…*
> *S3: [Student3 just starts to speak at the same time with the teacher]…a battery has no more…*
> *T: …voltage.*
> ***S2: So, do they go back there?***
> *T: The voltage causes it that the electrons start to move in a conductor.*
> ***S4: So, the electrons do not run out but the voltage run out…(?)***
> *T: It must be then… you can charge, for instance, a battery and get them (bulbs) to burn.*
> *T: [The teacher changes the topic.] In the beginning I asked, what is difference between a battery and a wall socket?*

This is a similar case as the previous example, where the teacher is confronted with a challenge as to whether and how to respond to the diverse inputs contributed by the students.

Example 5: During this lesson, where the next excerpt is taken from, the whole class is voting on and debating the different answers suggested by the students. The students (individually and then in groups of two) have already made their different predictions regarding the calculation of the area of a garden while using a special area unit given to them.

> T: *so the last one left* [answer]
> T: *9.8 at your opinion, Nassim?*
> S: *it's wrong*
> T: *Why ?*

S : *well because first it's too small*
T : *Why ? / So Nassim you think it's wrong/ why? /*
S : *Well no, actually I don't think that's wrong /*
T : *You don't think it's wrong / Anissa (asking another student)* »

The teacher elicits student answers about the information under discussion (if 9.8 is considered to be a wrong or a right answer). The teacher also seeks to engage other students in the discussion by asking for the opinion as to whether this seems like a reasonable value. The challenge here lies on the dilemma of the teacher whether to insist on probing the student to explain his/her reasoning or to move on to get the opinions of other students too.

Challenge 2

There is a challenge for the teacher on how much guiding s/he should be providing in cases that s/he recognizes that students have a particular difficulty. One of teacher's options in such cases could be to let students explore and learn from their own mistakes. This could turn out to be useful for certain occasions and certain students. However, this is not always the case. It might end up as very unproductive experience for students.

The following examples illustrate better the second aspect of this particular challenge, which relates to the effort of the teacher to steer the discussion in an effective manner.

Example 1: In the following excerpt, students present an experimental design that would help them investigate variables that might influence sinking or floating of an orange in water.

S We've decided to test the amount of water displaced by the orange and measuring it at 400, when you replace the orange.
T Okay, what is that going to teach you, what is that going to help you with?
S It's going to tell us whether the orange displaces more water, its own weight, less weight, or its own weight or the same.
T Okay and what will that tell you?
S Whether it's positive, negative or neutrally buoyant.
T Okay, cool and how will that help you know whether or not it's going to float or sink?
S Because if it's positively buoyant it will float and it will be because [inaudible].
T That's really interesting, keep going, I'm not judging you but that's just really interesting.

In this instance, the students did not seem to have a plan as to how to determine the weight of water displaced by the orange and it appears that the teacher had noticed this from the outset, but chose not to explicitly address this. Thus, rather than helping students appreciate that their plan required further development, he selected to challenge students on their

justification for their idea rather than guide them to successfully complete their method. The crux here is whether the teacher believes that students struggling to work out the pragmatics is justified in the time frames that students have to complete inquiries and it might be, for some students, that direction is needed to ensure the method would be successful. In such cases, the teacher could ask guiding questions to help students articulate beforehand their plan for how to conduct an experiment and how to interpret the possible results (e.g. what type of apparatus would you use for this?). The idea would be to guide the students to detect flaws in their experimental plan and revise it accordingly. Selecting not to intervene in such cases, leaving the students proceed with a problematic plan, is another dilemma the teacher has to deal with on time and, try to make the best pedagogical choice.

Example 2: In the next case, students have planned how to study Ohm's law. They are taking measurements, but they have problems with the wiring of meters. The teacher probably thinks that there are some problems with students' measurements, but stops guiding and goes to another student group. Students are left in an uncertain state. The teacher comes back later a couple of times and the issue is eventually resolved. However, the student pair loses working time of about 10 minutes. Probably, the teacher moves from group to group as she wants to ensure that all groups are on task.

> *T: Do you remember now, what [voltage] have you supplied to the resistor? How much was the voltage when you measured using this [voltage meter]? How much did you supply from here [voltage source]?*
> *S1: Yeah, it was six.*
> *T: And anyway, it showed eight?*
> *S1: Yeah.*
> *S2: Yeah.*
> *T: I see. [The teacher leaves.]*

This is another example of a case that the teacher faces a dilemma as to whether to allow students to struggle with a problematic situation or offer more guidance. At some point the issue will be solved, but students could lose much time and the question is whether this is compensated for by the process of allowing students time to work around the problem and come up with a solution themselves. This dilemma presses for a pedagogical choice in real time, which is certainly a challenging task.

Example 3: Students have got peer feedback about their plans confirming Ohm's law. They seem to not understand wiring of voltage and current meters based on the teachers' guidance. Students have difficulties also with electrical connections in series or/and parallel. Thus, how to go about connecting a current meter is an important issue.

> *T: And does the place matter where the current meter is?*
> *S1: Like this way.*
> *S2: I cannot make out from that.*
> *T: Erhm, it doesn't matter.*
> *S1: Eh?*
> *T: Does it matter where the current meter is located in the electric circuit?*

S1: Maybe not.
T: As long as it is somewhere.

The teacher seems not to be paying much attention to the connection of the current meter in the circuit or the underlying reasoning. Again, this might be a strategic choice to let students experiment and see if they can manage to set up the arrangement properly but there is still a challenge whether this is the optimum choice to make in the particular context in each case.

Example 4: The following excerpt is from a session which is part of a sequence of lessons where the class explored, trialed, investigated, designed and tested for a biscuit that meets specified criteria (e.g., taste, health, cost etc.).

> *T: Maybe, I mean, it all came out of the same pot, but maybe their tea might have been hotter than their [another group's] tea. What else might have been different? I'm thinking about the size of something, Christiana?*

Teachers often have a clearly specified route for how the discussion could/should evolve. In some cases, it might be difficult for them to allow deviations from this route, depending on how the discussion actually evolves. It is possible that the teacher might end up closing her line of questioning so that an initially open question ('What else might have been different?'), which could elicit useful contributions by the students, was followed by a closed question that essentially made any such contribution highly improbable (e.g., I'm thinking about the size of something). It is important to note that during the interview session, this teacher stated that she sometimes deliberately wanted to guide students when she felt they had become stuck on something:

> *"I guess the biggest downside probably is the 'what do you do when you get stuck' thing. The students who were really stuck or students who were sort of going in a direction which I didn't think was going to solve the problem, I think, one way or another, whether I sort of did it through questioning, making them think about it, but one way or another, I got them onto the track that I had in my mind".*

Example 5: The following example provides another illustration of the challenge faced by the teacher, with respect to deciding which way to use students' contributions and that decision is often restricted from other parameters like time constrains. In this case, the teacher is confronted with a question to which s/he could respond either with probing students to elaborate more on their thinking or to guide them to the correct idea.

S9: Can we set these this way and attach to this lamp?
T: Erm -, but then you only continue the wire. Now, you have a short circuit. You have a short circuit. You connect with the wire the terminals to each other.
S9: But how do we get like ... when there was a drawing [refers to students' worksheets] where these were like this?
T: ***Yes, was that drawing correct?***

S10: Well, no.
T: Yes. Well, if it does not light. It was a short circuit.
S10: Well, then we cannot try that, if ...
T: You feel here that the battery will warm up if it short circuits for a long time.

Challenge 3:

The discourse that unfolds in the classroom is complicated by the possibility that the inputs contributed by the individual students or the teacher, could be resting on tacit assumptions, not necessarily shared by all members of that specific learning community. This might influence how the classroom discourse evolves; thus there is a challenge associated with detecting such cases which can occur in classroom discourse.

Example 1: There might be a variation associated with the extent to which these instances surface in the discussion in a profound manner. This could have an effect on the demands that are placed on the teacher in terms of detecting and managing such instances. In the following excerpt from a classroom dialogue, a crucial misconception that gravitational force is not an instance of interaction between masses but an inherent property of a single object (e.g. the earth as a whole – including the atmosphere – or the ground) is identified.

T: Hmm. If I left it, it falls… Why? Is it because of the ground?
S3: The gravity.
T: What is gravity?
S3: The force towards the Earth's centre.
T: Towards the Earth's centre. So the object falls because of the gravity?
S3: Yes
T: Where does the gravity come from? From the ground?
S4: From the Earth's centre.
T: So if we somehow remove Earth's centre won't there be gravity?
S1: From the Earth in general
T: From Earth in general. Ok…

This subtle point, which was not addressed during the discussion, despite the opportunities for doing so, came into play a little bit later during the same dialogue as illustrated below, when the teacher suggested ignoring the air in order to avoid discussion about its resistance when the ball is falling.

T: Let us assume that there is no air. In these environments we will consider that there is vacuum.
S: If it was vacuum, the ball wouldn't fall.

In the excerpt above, we can identify a student's misconception that gravity needs air as a mediator of the interaction between objects. In this case, it is more challenging for the teacher to appreciate the students' point of view (gravity as a feature of the earth - as an object - rather than an effect of the interaction between the large mass of the earth and other objects – e.g. the objects placed on the table in the system of interest) because it is not explicitly expressed. Thus, one likely interpretation for why she had avoided clarifying

this issue is that she simply did not recognize students' interpretation of gravity as a property of the Earth. During the interview session that we had conducted with the teacher after the implementation, having her watch the particular video-excerpt, she stated that she realized that more students didn't have a valid understanding of the concept of gravity. Nevertheless, even then it was hard for her to understand how students conceptualized gravity. Overall, this example illustrates how challenging it is for the teachers to identify hidden assumptions that students have about particular concepts and respond in real time, in a way that could help them overcome their misconceptions.

Example 2: In the next example the theme under discussion is the relation between temperature evolution of water (heated to boil) with time, heat transfer between the heater and water and the interpretation of the phenomenon at a microscopic level (at the level of particles).

> T: *so it's your turn* [pointing to another group]
> St3: *well we think that / it [temperature] will get higher because as the heat particles they will accumulate / first they will be at a certain temperature and then they will accumulate / so it will take much more place so it will accumulate / but on another hand it will evaporate so it will continue to raise it means more heat / (inaudible)*
> T : *Ok / so you think that there's heat particles that will accumulate inside water and they will make temperature rise / so is that it?*
> St3: *yes*
> T : *So that's why the temperature will rise? /*
> St3: *it will evaporate / (inaudible) / since it will rise till a certain temperature unless to have / (inaudible) /*
> St4: *it will not rise in one shot / (inaudible)/*
> T : *Ok you think too that temperature will rise faster and that water will evaporate faster and that the straight line drawn on the graph / the temperature rises faster on this graph than on others /*
> St4: *Yes*
> T: *What do you think about all that / (speaking to another group)? …*

The students talk about the reasons why the liquid water temperature continues to rise during boiling, by means of energy transfer through heating. Students seem to understand that there is heat transfer from the heater to the water. But heat transfer is described by the students as, a "heat particle" that is exchanged between the heater and the water (they say later with the room too). It's akin to materialization of a physical concept. The teacher during this theme, besides exploring students' ideas, didn't seem to intervene to reduce the gap between students' conceptions and the learning goals. The teacher continues to try to encourage students to continue to debate on the issue by asking other groups to give their opinion, even after she asked students to formulate a counter-argument.

In the interview conducted with the teacher after the lesson, the teacher mentioned that she had noticed students' conception (that a drop of water will store two types of particles

water and energy particles so it can evaporate and then water will leave part of the energy-particles (in the room) and keep some of these particles) and found it an "interesting idea" that (the energy) could be quantified/materialized. But she added that she did not feel comfortable or well-prepared to engage in a discussion about the physics of the specific situation at that particular time and she thought that it would be better not to probe this further. She added that especially with on the fly interactions, press for immediate actions on the part of the teacher and do not allow revisiting background knowledge (thermodynamics) to decide how to deal with the ideas offered by the students. This example reveals how challenging it is for the teacher to respond to students' ideas as they emerge in the discussion.

Example 3: In some cases, the students and the teacher might be drawing on the same technical terms without necessarily ascribing to them the same meaning. In a later part of the dialogue of the previous example (Example 2), the teacher seems to be considering that the student shares with the teacher the same meaning of "heat" (as an energy transfer from the heater to water). Students could hold different conceptions regarding heat. However, the teacher doesn't allow the students to express themselves in detail regarding what they understand by heat and the teacher straight away says it's energy ("*yes because of heat / what are we going to call this? / The energy / So that it has been given by the heater")* and then passes on to another group.

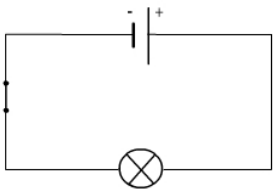
Consider the following excerpt from the whole class discussion, where the teacher tries to help students define the notion of the "environment" for an object in the context of dynamics, discussing about a small ball that she holds in her hand:

> *T: Are plants, sun, the atmosphere part of the environment for this object?*
> *S: No.*
> *T: Why not?*
> *S: Because we are not interested in whatever is outside.*
> *T: Right…*
> *[…]*
> *T: What is the difference between ground and Earth?*
> *S1: The Earth in general.*
> *T: The Earth in general and the ground in particular?*
> *S1: Yes, because the ground is in its (the object's) environment.*
> *T: The Earth isn't in its environment?*
> *S1: The Earth is outside.*
> *S2: It (Earth) is natural environment.*

It seems that students hold the misconception that the environment for an object is what is spatially close to the object. This students' conceptualization does not seem to be realized by the teacher. Consequently, both the students and the teacher use the same term (environment) but in a different sense. This seems to have a negative influence on students' understanding later in the dialogue when they classify Earth as not a part of the object's environment (in terms of physics) but only as an object of the natural environment. In this case, the miscommunication lies exactly on the term under discussion and not on a relevant

concept that came up during the discussion, like gravity in the previous example. However, it seems that in this case it is also challenging for the teacher to identify and address the deviation in students' conceptualization of a term as compared to the one she would like them to have.

Example 4: In the following excerpt, a student asks a few times about the representation of a closed circuit. The problem seems to relate to the gap in the symbol of the battery. Probably, the student wonders how electricity can flow over the gap. The teacher gives an explanation directly without probing the student's (mis)understanding.



S: Is that [refers to the picture] closed?
T: This is closed now, because there is a route for electrons from here to there [shows the route].
S: So, that works?
T: That works. Electric current flows now. If it is an open circuit, it is open at some point. [The teacher continues her explanation.]

Students' contributions to the dialogue suggest the presence of a misconception which is not addressed by the teacher. An interview reveals afterwards that the teacher did not think that the student had such problem with the representation. The teacher considered that she could have asked more probing questions to understand the student's thinking. However, she said that one reason why she did not probe was that it was the end of the lesson and the time was running out. This example illustrates that it could be difficult for the teacher to realize possible misconceptions that might underlie students' statements or questions.

5. Assessment method: Structured Assessment Dialogue

5.1 Information on how structured assessment dialogue was incorporated in the teaching intervention

The Structured Assessment Dialogue (SAD) has been designed to be able to combine a formative and a summative use of an assessment process. The implementation of this assessment method conformed to certain specifications, which were formulated at the level of the project with the intent to ensure a robust structure that could allow studying the features of structured assessment dialogue, as a formative assessment method, in operation. These specifications were as follows:

- The LWGs had to develop teaching and learning materials targeting at least one of the six competences under investigation.
- The teaching plan includes a ritualized conversation (5 minutes) between one student (the student in focus) and a teacher based on the student's preparation and a template filled in by the teacher mirroring the demands for the competence to be assessed. This is followed by a peer-feedback phase where the student in focus gets feedback on this or her presentation and performance by a group of feedback students. During these two interactions, the rest of the class observe and reflect on their own understanding. In the end, all students write down reflections about their perceived level of competence and future learning path using a reflection tool.
- Data is produced at all stages of the process.

In total, in this third round, this assessment method was implemented in nine cases, each comprised of several implementations, in a total of 37 SAD sessions by approximately 16 teachers. Table 61 summarises the parameters associated with each of these nine implementations in terms of the targeted competence, educational level, number of dialogues, subject and specific topic in which the dialogues were situated.

Table 61

Implementations based on Structured Assessment Dialogue

	Competence	Educational level	Number of dialogues	Subject	Topic	Country
1	Investigation/ Modelling/ Argumentation	Upper secondary	8	Physics	Kinematics, Energy forms	Denmark
2	Investigation/ Modelling/ Argumentation	Upper secondary	2	Biology	Ions/nerve signals	Denmark
3	Design/ Innovation	Upper secondary	4	Technology	Electrical circuits	Denmark
4	Investigation/ Modelling	Lower secondary	1	Science	Electrical circuits	Denmark
5	Problem Solving/ Argumentation	Upper secondary	5	Mathematics	Programming	Denmark
6	Investigation	Upper secondary	2	Physics	Density & Archimedes Principle	United Kingdom
7	Investigation	Upper secondary	2	Biology	Membrane permeability	United Kingdom
8	Investigation	Lower secondary	8	Physics	Electromagne tism	Finland
9	Problem Solving	Upper secondary	4	Mathematics	Geometry	Finland

5.2 Research Data

In accordance with the specifications that were formulated for the implementation of this assessment method, the episodes of Structured Assessment Dialogue were video recorded. In this episode, the students have a specific role to employ. Figure 13 illustrates the research method and shows the research data. For each implementation the following raw data were produced:

- A filled-in teacher preparation template
- Video (or audio) clip of the 5 minute dialogue
- Audio clip of the 5 minute peer feedback session
- Filled-in Student Self Reflection Forms (from all students)

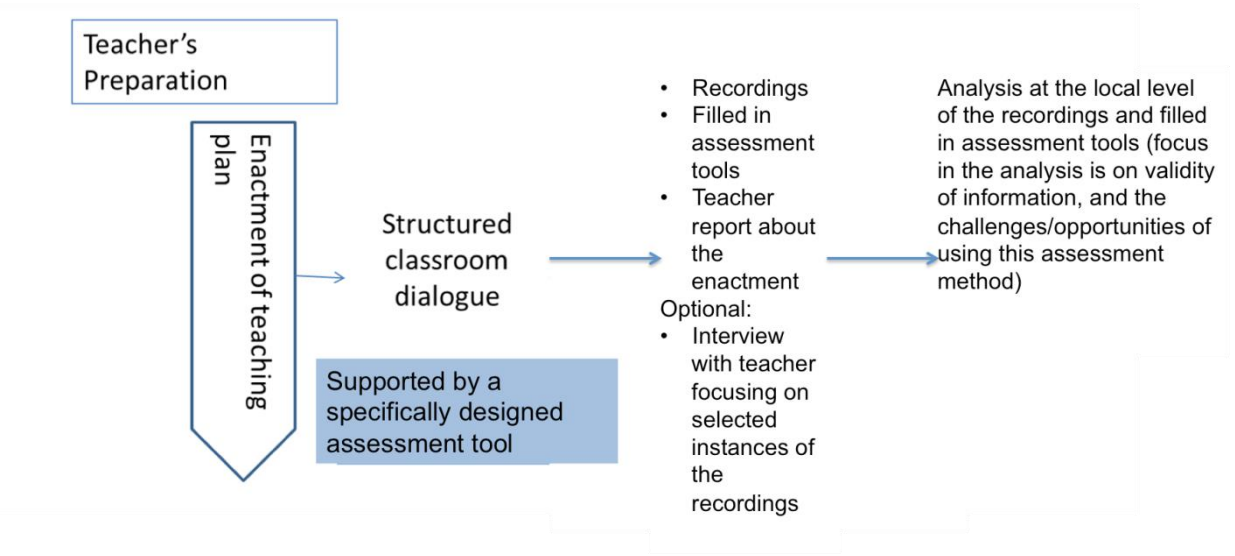


Figure 13. Illustration of the Research Design

5.3 Data analysis

All the video and audio clips and the completed templates were collected. The researchers undertook to process and code the data using specially designed network analysis tool for the videos in order to be able to categorize the dialogues. The coded data was analyzed with the intent to detect and document emerging patterns about aspects of the SAD as an assessment method.

5.4 Results

This section presents the results that emerged from the synthesis of the data reported from the various implementations. It is structured in two sub-sections. The first pertains to the analysis of the episodes of SAD whereas the second reports insights into the teachers' perceptions about various aspects of the implementation of this assessment method.

5.4.1 Analysis of teacher-students dialogues

The data collected from three countries (Denmark, UK, Finland) that implemented SAD as an assessment method can be analyzed at two levels. The first level offers descriptive information about the duration of the various types of contributions that appear in the teacher-student dialogue in an episode of a SAD. The second level of analysis includes patterns of the teacher-student dialogue. One issue is to categorize the dialogues according to their ability to cover all aspects of the assessed competence in breadth and depth. This pertains to one of the key ideas of the SAD: To make the demands clear for the students and to give them the possibility to assess their own competences in relation to the standards set by the teacher-student dialogue. Is this approach, a dialogue has quality when it gives students possibility for formatively and summatively assessing their own competence. This analysis is based on a network analysis of the coded video clips. This analysis is in progress and not finished at the finishing of D5.11. Another issue for analysis of the dialogue is how students are led to high order contributions and this will be pursued in the following. This could be seen as an indication of a high quality dialogue, although the quality of the dialogue necessarily must be seen in relation to the purpose of the dialogue. For analyzing the

collected data, teachers' and students' contributions during the dialogues were coded using the scheme shown on Table 62.

Table 62

Codes for teacher's and students' contributions

Teacher's contributions		**Student's contributions**	
Category	**Brief description**	**Category**	**Brief description**
Precise valuing	The point is that the teacher is precise and puts value to what is said.	Students' question/ not understanding	It is quite a broad code basically covering parts of the dialogue when the student isn't sure what to answer, hesitating, or asking the teacher clarifying questions
General Evaluation	Mean as a possible counterpart to precise valuing - but this code is for general praise/criticism.	Lower Order Answers & Statements	Covers the knowledge and comprehension steps
Focus	Focus can be seen as an emphasis on the set teaching goals, where uptake can go out on a tangent.	Higher Order Answers & Statements	Covers application, analysis, synthesis and evaluation
Invitation	The teacher invites the students to engage in the dialogue.		
Uptake	"Incorporating students' responses into the next question, thus getting the students to reflect further about what they said, and integrating the answer into the dialogue[...]" (Quistgaard, 2014b)		
Summarizing	The teacher rephrasing, clarifying or summing up what the student is saying without valuing or correcting it		
Lower Order Question	Covers the knowledge and comprehension steps		
Higher Order Question	Covers application, analysis, synthesis and evaluation		

Level 1: Descriptive information about the various types of contribution that appeared in Structured Assessment Dialogue

As shown in Figure 14, the teacher posed higher order questions for most of the time in episode 13. Also, there were instances of uptake that opened the discussion. An uptake seems to provide quality in the discussion, because it broadens students' thinking. However, as shown in Figure 15, this seeming quality of episode 13, on the part of the teacher, did not lead to a corresponding productive dialogue on the part of the students. In particular, the appropriate background for productive dialogue that was formulated in episode 13 was not accompanied by a productive response on the part of the students. On the other hand, in episode 20 which is also one of the most productive episodes of SAD from the teacher's perspective, the students take advantage the appropriate background, and retain the high level of the discussion, posing high order answers and statements.

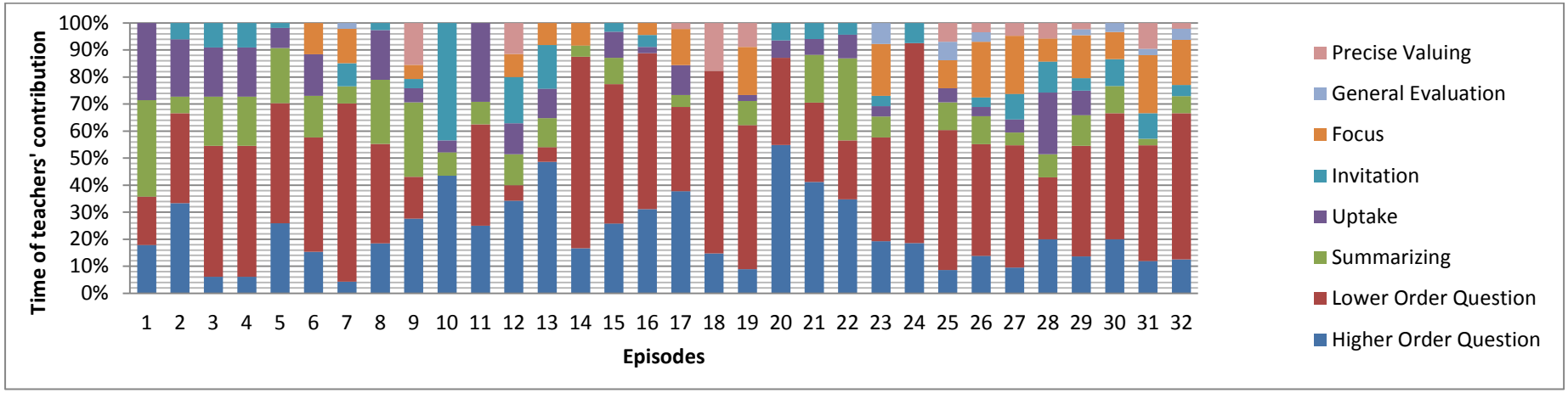


Figure 14. Teacher's contributions from all the episodes of *Structured Assessment Dialogue* in Finland and Denmark

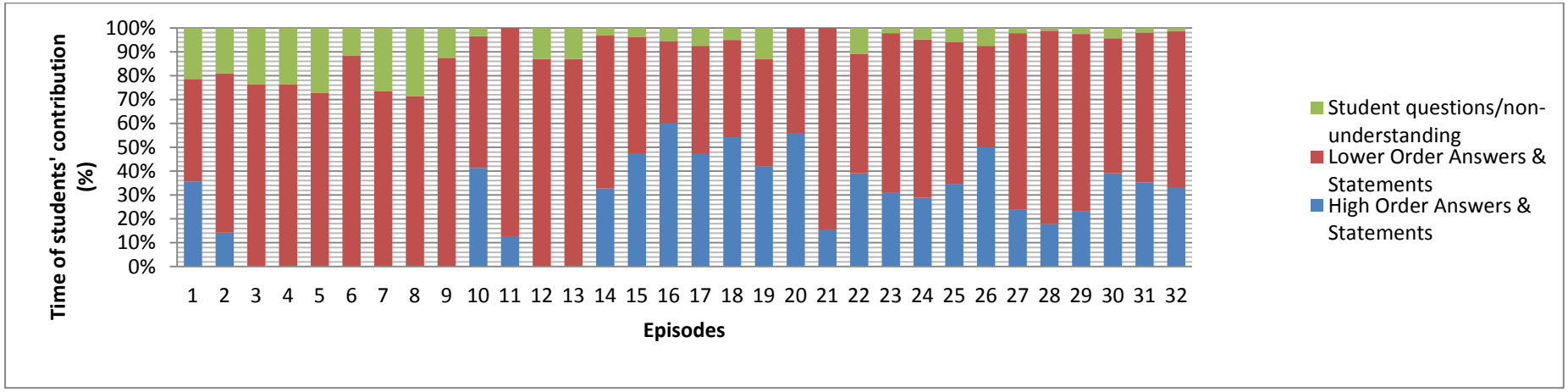


Figure 15. Students' contributions from all the episodes of *structured assessment dialogue* in Finland and Denmark

It is a bit surprising that there doesn't seem to be a correlation between the order of teacher questions and of student responses. For instance, in episodes 3, 4 and 7, there are few higher order questions and no higher order answers or statements, whereas in episodes 18, 19, 25 and 26, correspondingly low frequencies of higher order questions are seen together wither higher frequencies of higher order answers/statements. Seemingly, the frequency of higher and lower order questions/answers are not in itself capturing the essence of the dialogues. Therefore, we have tried to invent a different approach able to take into account the complexity of the dialogues, showing the relations between the dialogical elements. This method will be illustrated in the next section, the level 2 analysis.

Level 2: Patterns of teacher and student contributions in the dialogues

Figures 16, 17, 19 and 19 represent the contributions (as listed in table 62) made by the teacher and the students during an episode of SAD. It plots the sequence of the various types of contribution made during the episode as a function of time. A pattern that seems to underlie the data is that, when the teacher challenges the students by means of high order questions (code 7), the students respond with high order answers (code 7.5). This gives the opportunity to the teacher to broaden the conversation with an uptake (code 8).

Table 63

Codes that are presented in timelines.

Codes for contributions:

Teacher:	Student:
1. Invitation	2,5. Lower Order Answer
2. Lower Order Question	7,5. Higher Order Answer
3. Precise Correction	
4. Precise valuing	
5. Summarizing	
6. Focus	
7. High Order Question	
8. Uptake	

The relation between the teacher and the students' contributions is also evidenced by the fact that lower order questions from the teacher are usually followed by lower order answers from the students. In other words, the data seem to provide support for the claim that students' contribution to the dialogue are significantly influenced by the teachers' own contributions, which, in turn, implies that the facility of the teacher to productively steer the discussion seems to be decisive.

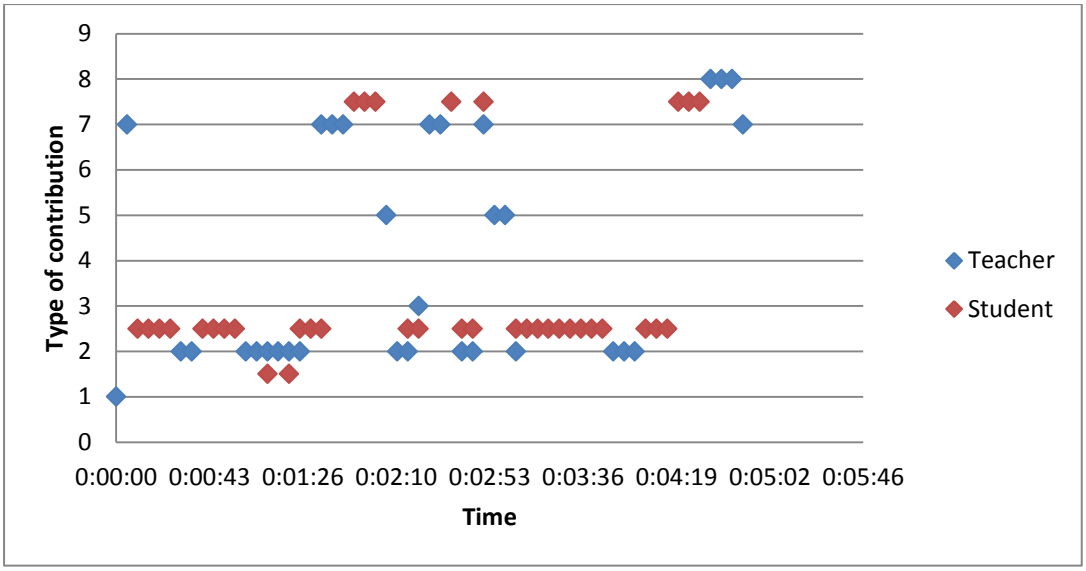


Figure 16. Timeline the episode 2 of *structured assessment dialogue* in the case of Denmark**.**

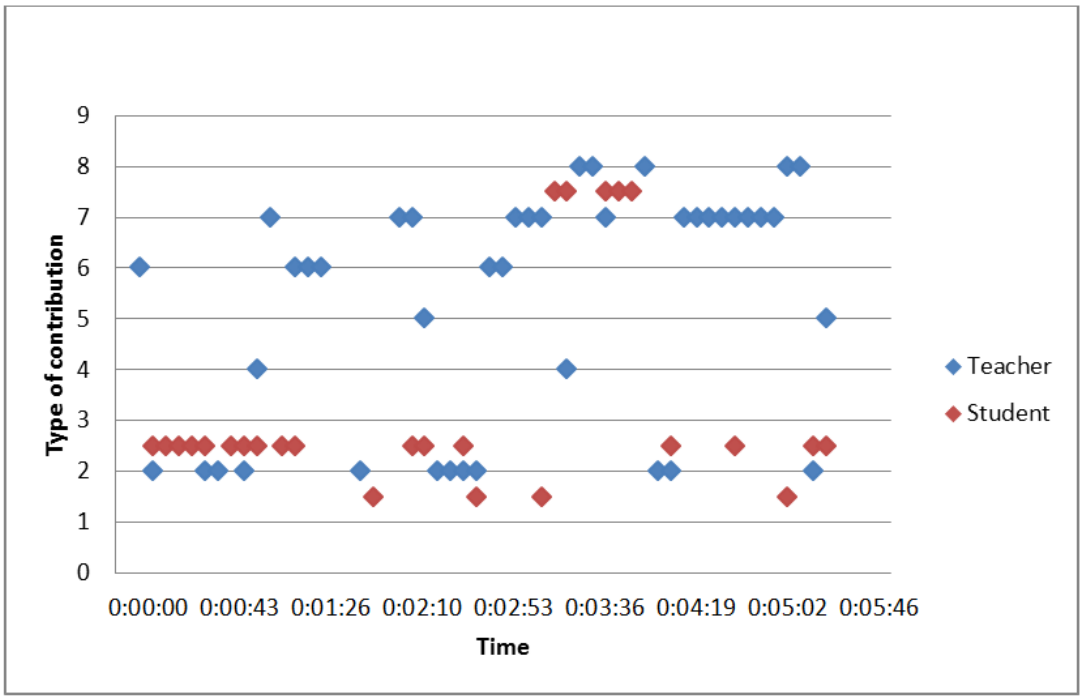


Figure 17. Timeline of the episode 4 of *structured assessment dialogue* in the case of Denmark.

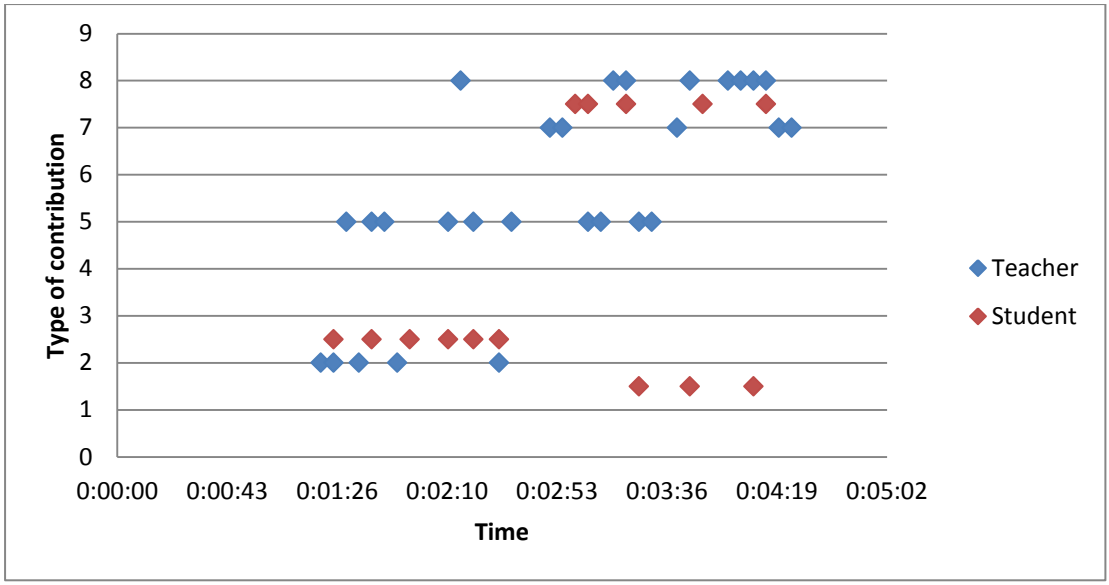


Figure 18. Timeline of the episode 1 of *structured assessment dialogue* in the case of Finland.

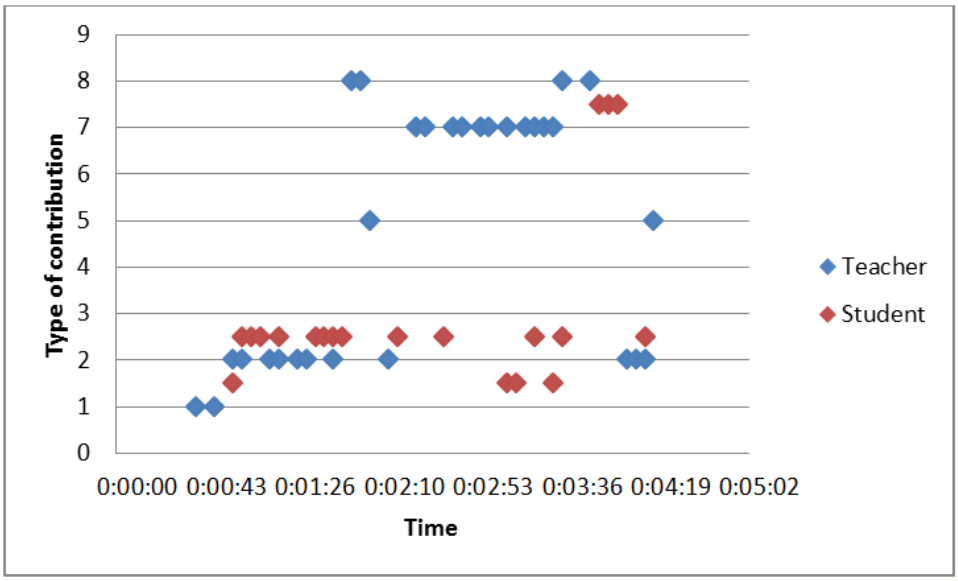


Figure 19. Timeline of the episode 2 of structured assessment dialogue in the case of Finland.

In addition, the graphs presented above suggest that certain teachers' contributions (e.g., precise correction, precise valuing, summarizing and focus), are not followed by higher order contributions from the students. But these teacher contributions can have purposes pertaining to the first issue of the dialogue, namely the mapping of the competence demands at hand. This implies that there are certain types of contribution by the teacher that are not likely to enhance the intellectual level of the dialogue but rather establishing the foundation for the students' self-assessment.

The data from a SAD episode in Finland (Figure 20), indicated another pattern where, a lower order discussion (lower order questions and answers) was followed by the teacher proceeding to summarize the discussion while, in cases of a high order discussion, the teacher usually responded by means of uptake.

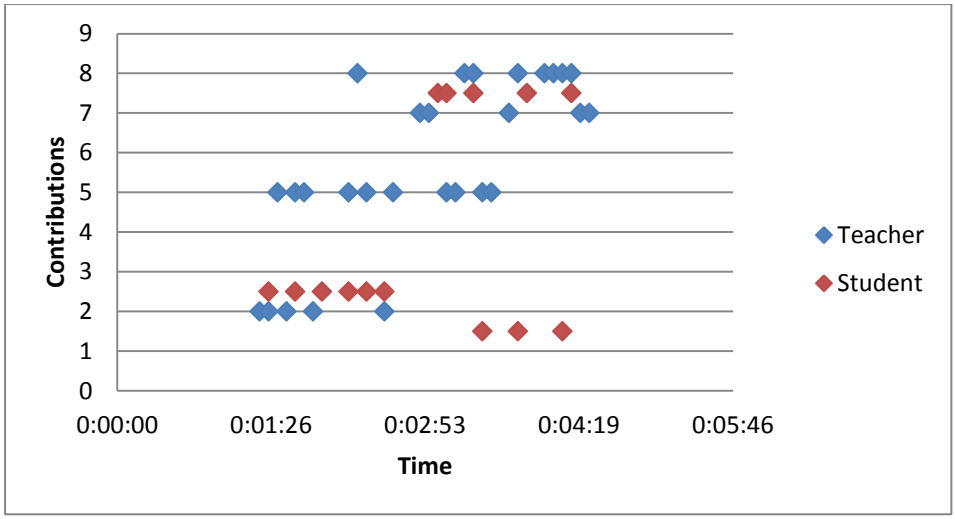


Figure 20. Timeline of an episode of structured assessment dialogue in the case of Finland

To supplement the previous approach, we used an iterative mixed methods methodology to create and interpret network maps of student-teacher dialogues. The iterative nature of our methods are imperative to creating meaningful network representations. It is not trivial to relate the structural patterns that emerge in such networks to dialogical learning concepts. The aim was to capture both dynamic and structural elements of the dialogue as they appeared in the video recordings. To achieve this, we coded the actions of both student and teacher in five-second intervals. The results of the coding process were spreadsheets coding the video dialogue for the teacher's and the student's content, dialogical action, and gesture actions in 5-second intervals. To create networks from these spreadsheets, we concatenated codes in a 5-second interval to a single code. The result of this was timeline of codes describing how the dialogue progressed in time. We then represented the concatenated codes in a network. Figure 21 shows two such networks.

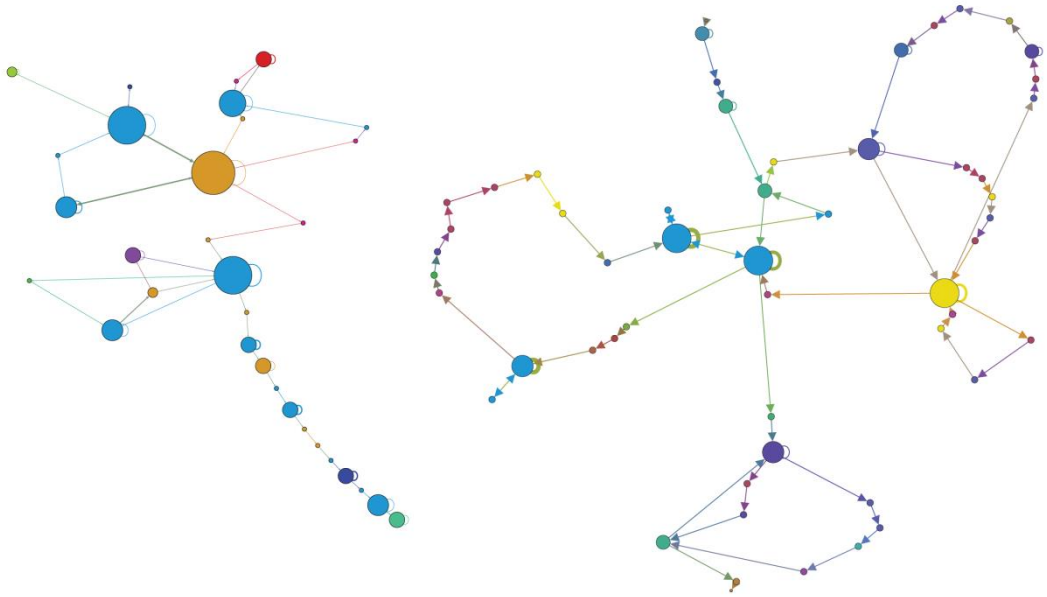


Figure 21 Example dialogue networks

The two networks depicted in Figure 21 seem to have very different structures. For example, the network on the left seems simpler and more linear than the one on the right. In this representation, we have used colours to represent different aspects of the codes. For example, nodes that represent student speech actions are yellow, while nodes representing teacher actions are blue. Nodes that represent speech actions of both student and teacher are green. Comparing these two networks, it is clear that the teacher is most prominently represented in the network on the left, while the network on the right shows a more equal distribution of speech acts.

This is where we are by the delivering of D5.11. We will in the coming months categorize the networks and assign each category to a specific quality of the dialogue. Each SAD enactment will then be represented by a network category and we will look for relations between the quality of the dialogue, as represented by the network category, and the usefulness of the dialogue as measured via the students' self-reflections.

5.4.2 Analysis of student self-reflections

In the final part of a SAD session, all students were given a self-reflection template where they were asked to reflect about their own current level, what questions they would and would not have been able to answer and what their next learning steps should be. Furthermore, they were asked about how well the student-teacher dialogue and the feedback session, respectively, helped them understand their own level and determine their next learning steps.

In Figure 22 and Figure 23 student answers to the latter of these questions can be seen for implementations in Denmark and the UK. Two different reflection tools were used in the two countries, meaning that in Denmark, students answered on a 7 point Likert scale

whereas in the UK, students answered on a continuous scale from 0 to 100 % which was then interpreted as a 5 point Likert scale by researchers. Furthermore, students in the UK only answered the questions about their current level and not the ones about their next learning steps.

In the figures, the mean of the student answers can be seen for each question. The reflection tool was not used in all implementations, meaning that the figures represent data from a total of 14 SAD sessions.

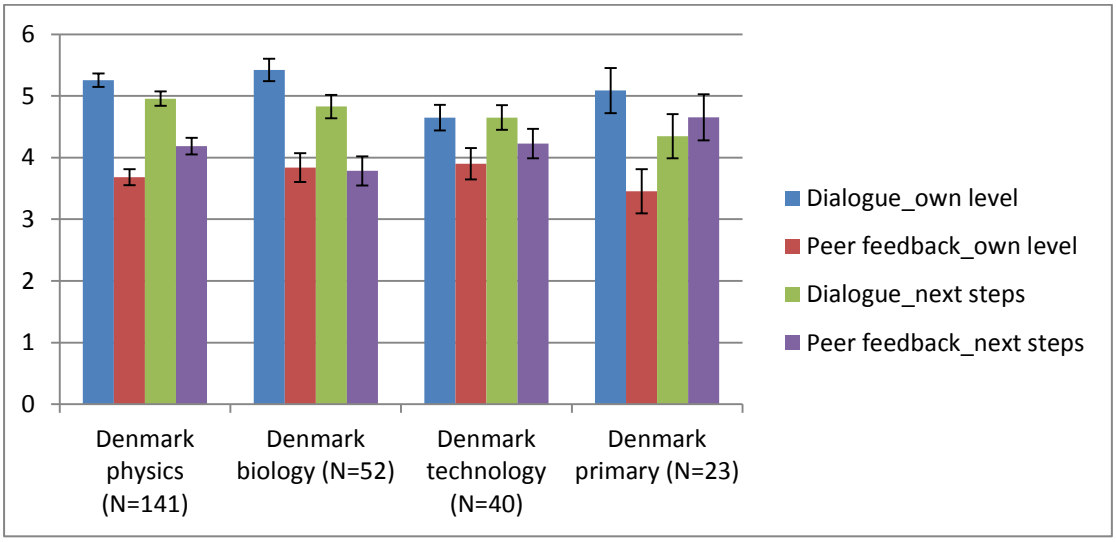


Figure 22 Student self reflection from Denmark. Error bars represent 1 standard error.

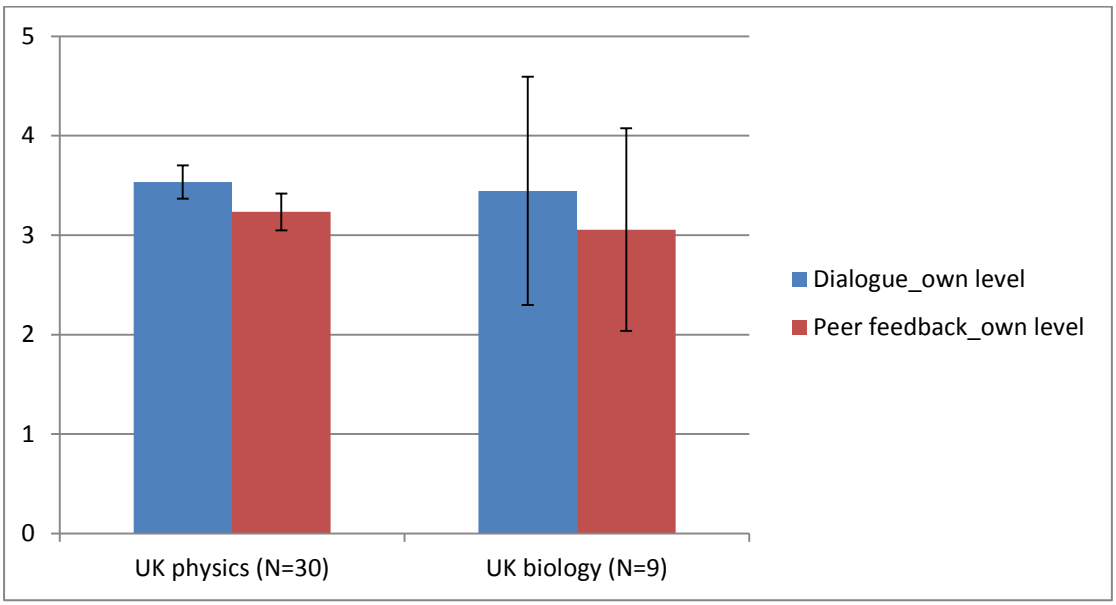


Figure 23 Student self reflections from the UK. Error bars represent 1 standard error.

As it can be seen from the figures, students generally found the teacher-student dialogue to be more helpful than the feedback session in their own self-assessment. This is especially prevalent with regards to determining their own current level of attainment, which is

probably not very surprising since one of the main aim of the dialogue is to make the demands clear to the students and the formative aspects of the assessment method are often mostly present in the feedback session.

These results are well aligned with a general understanding among the teachers that it was difficult for the students to provide useful and productive feedback. However, that the dialogue phase of the SAD method was often found to be more useful by the students does not mean that the feedback phase can be omitted, since a lot of the formative elements of the method can be found here.

Teachers' perception of the method

Generally speaking, teachers from Denmark and the UK found the implementations to be well aligned with their usual teaching practices, whereas the method was more unfamiliar in a Finnish context.

Some Danish teachers said that the 5-minute duration of the student-teacher dialogue was a bit short to thoroughly cover all aspects of the learning progression they had intended for the session. This meant that they had to be careful in not being overly ambitious when choosing what to assess. For instance, one teacher used SAD in a teaching unit about electrical circuits and found that he had to reduce the complexity of the circuits the students had made in order to be able to discuss all components with the focus students. On the other hand, some teachers stressed that the clear rules and short duration of the student-teacher dialogue helped to make the rather unfamiliar situation feel less intimidating for the focus student.

Generally, teachers from all three countries seem to agree that the peer feedback phase proved to be challenging for the feedback students. It requires practice to learn to give useful feedback and in some cases, the students needed quite a lot of scaffolding from their teacher in order to give feedback. There could be different reasons for this. Some Danish teachers said that the learning progressions were difficult for students, especially since they aimed to assess competences rather than content knowledge. This meant that even though students might be able to determine whether what was said by the focus student was right or wrong that were not able to relate it to the teacher's demands or give suggestions for next learning steps. Furthermore, some Danish teachers pointed to a classroom culture where students perceive feedback as criticism rather than help and are thus reluctant to point to places in the dialogue that could be improved.

6. Teacher Self-Reporting tools

The teacher's self-reporting tool consisted of a set of various Likert scale items and four open-ended questions about the teacher's perceptions concerning (a) the process of enacting the specific assessment method in their context (e.g. time requirements), (b) how certain features of the assessment method played out, and (c) how students engaged with each assessment method (written feedback, peer-feedback, on the fly interactions). These items asked teachers to indicate the extent of their agreement/disagreement with given statements. The first level in this scale implies absolute disagreement, the highest level implies absolute agreement and the intermediate states indicate positions within these two extremes. The second part contains four open-ended questions. Teachers were asked to reflect on challenges, opportunities, strengths or weaknesses of the assessment method. Furthermore, they were asked to suggest potentially useful resources that could help addressing these challenges.

Tables 64, 65, 66 and 67 present the results from the Likert-scale items for all the countries in all implementations of all the three rounds for four assessment methods, namely written feedback, Peer(self) feedback, interactions on the fly and structured assessment dialogue. One point of caution that has to be made from the outset is that the number of teachers who completed the tool is limited, prohibiting reliable inferences about underlying patterns. This limitation notwithstanding, in what follows we seek to briefly discuss the results that have emerged.

6.1 Written feedback method

One finding is that the teachers who enacted this method found it rather easy, or easy to implement (12/14) (sub-question 1.1). Also they found it rather easy to interpret students' difficulties and diagnose needs (10/14) (sub-question 1.2). Furthermore, most of the teachers stated that they found it easy to provide descriptive comments that could be helpful to the students (9/14) (sub-question 1.3).

In addition, teachers' opinions about the time needed for implementing the assessment method of written feedback varied considerably (sub-question 1.4). One possible reason for those who believed that written feedback is a time consuming method (6/14) is the inconsistency with their usual teaching approach (sub-question 1.7) or typical teaching practice in their country (sub-question 1.8).

Most of the teachers (13/14) seemed to agree that written feedback was beneficial to the students (sub-question 1.5) because the students: a) used to work with individualized descriptive feedback comments (sub-question 2.1), b) were in a position to interpret the feedback comments (sub-question 2.2), c) were keen to engage with reflection on the comments they received from their teacher (sub-question 2.3) and d) took into account the feedback comments while revising their artefacts (sub-question 2.4).

The teachers also noticed that the written feedback method was beneficial to them as well, (13/14) (sub-question 1.6) because they could produce information that may inform their instructional decisions about their next steps in a specific teaching unit (sub-question 1.10).

Moreover, the majority of teachers believed that this type of formative assessment could potentially serve summative assessment purposes (12/14) (sub-question 1.9).

Table 64

Results from the closed questions

1. The assessment method "Grading & Written Comments":	Total			
	1	**2**	**3**	**4**
1.1. was easy to implement		2	6	6
1.2. was easy to interpret students' difficulties and needs from the artefacts they created		1	10	3
1.3. was easy to write descriptive comments that could be helpful to the students	1	4	5	4
1.4. took more time than I expected	4	2	4	4
1.5. was beneficial to the students		1	8	5
1.6. was beneficial to myself as the teacher		1	8	5
1.7. is consistent with my usual teaching approach in my class	1	7	5	1
1.8. is consistent with the typical teaching practice in my country	4	8	1	1
1.9. could potentially serve summative assessment purposes	2		7	5
1.10. could produce information that may inform my instructional decisions about my next steps in the unit		1	7	6
Students were:				
1.11. used to work with individualized descriptive feedback comments	5	2	5	2
1.12. in a position to interpret the feedback they were provided with		4	9	1
1.13. keen to engage with reflection on the comments they received from their teacher	1	3	6	4
1.14. keen to take into account the feedback comments in revising their artefacts		4	7	3

This paragraph focuses on the synthesis of the results of what the teachers in Cyprus, Denmark, Switzerland and Germany considered as the main strengths and main weaknesses of written feedback as an assessment method. Teachers noticed that the most important strength of this assessment method is the individualized feedback that it provides. The feedback is tailored to the needs of each student or each group of students. Consequently, the learning process is more productive and this is reflected on students' revised artefacts. The following quotes illustrate this:

> *Students were provided personal feedback based on the answers they gave and hence they were encouraged to keep studying.*
>
> (Cyprus)

> *Each student gets an individual feedback concerning his/her experimental plan and can work on his/her weaknesses*
>
> (Germany)

The main weaknesses of this assessment method referred to the time needed for the teacher to provide individual feedback and the often violated condition that the students will take seriously the teachers' comments. The following quotes exemplify these limitations:

> *Feedback should be written within a single day and given back to the students on the next day. I personally had only limited time available for writing comments, due to other commitments. It is worth noting that my class includes only 12 students. I believe that if it consisted of more students, this method would be even more difficult to implement.*
>
> (Cyprus)

> *Not all students take criticism from the teacher seriously*
>
> (Switzerland)

One opportunity for written feedback recognized by the teachers is that it could help them gain insights into their students' misconceptions and difficulties and therefore they could adjust their instruction to more effectively address their students' needs. From the perspective of students, this method is fruitful, because they can be more engaged in the learning process. The quotes below are indicative of this:

> *Feedback offers teachers the opportunity to gauge their students' state of scientific knowledge, assess their skills and identify their weaknesses.*
>
> (Cyprus)

> *The students are forced to reflect on concrete aspects in the assignment. The assessment becomes more a process, rather than a product, which is certainly beneficial for the student's* learning.
>
> (Denmark)

The challenges that derived from the teachers' experiences with this method focus on four main aspects: a) provision of systematic and understandable feedback, b) consideration of students' individual needs, c) alignment of feedback with the learning goals of each lesson and d) making written feedback a less time-consuming procedure. One suggestion that teachers provided in order to overcome these challenges includes the formulation of guidelines or tools that could potentially support teachers on how to give effective written comments. The following quote illustrates this:

> *Feedback in the form of a checkbox rubric (similar to the used evaluation sheet) with individual task part where students may reflect only - supported by peers from their group – on preselected parts of their plans and correct these. This would reduce the amount of time needed to provide feedback which would allow to use the method also in the daily teaching practice*
>
> (Germany)

Moreover, teachers should supplement written feedback with accompanying oral clarifications where useful. This could increase the probability for students to actually act on the feedback. The quote below shows this suggestion:

> *It would make sense to discuss the feedback with the individual student groups instead of just distributing it in a written form.*
>
> (Switzerland)

In addition, teachers need professional development in order to be able to provide written feedback comments to students. (Germany).

If the teachers had the opportunity to repeat the implementation of this assessment method, they mentioned that they would make certain changes to specific parts. First, they would provide students with fewer and simpler comments in order to decrease the cognitive load that is incurred to the students for understanding the substance of the feedback (e.g. "I would focus my feedback to include just an area in the assignment they made, e.g. the discussion", Switzerland). In addition, they would provide more time to students to revise their artefacts. In this way, the students would be able to reflect on their initial answers based on the given feedback. They would also provide students the assessment criteria in a written form.

6.2 Peer feedback method

The majority of the teachers found the assessment method of peer feedback easy to be implemented (18/31) (sub-question 1.1) and they reported that they do employ it in their teaching approach (7/9) (sub-question 1.4). Nevertheless, most of the teachers noted that it is a time consuming method (5/9) (sub-question 1.2) and they do not employ this method in their daily teaching practice (sub-question 1.4).

In addition, most of the teachers mentioned that this method could be easily employed in their teaching (7/9) (sub-question 1.3). Furthermore, 7 of the teachers stated that this method is not consistent with the typical teaching practice in their country (sub-question 1.5).

Moreover, the perceptions of the teachers about the potential use of formative assessment for summative purposes varied. Specifically, some teachers believed that peer assessment could not potentially serve summative assessment purposes whereas some others held the opposite opinion (sub-question 1.6). Additionally, all the teachers stated that this method could provide information that might be useful for the instructional design of the next steps in the teaching unit (sub-question 1.7).

Table 65

Results from the closed questions

1. The assessment method "Self &Peer feedback":	**Total**			
	1	**2**	**3**	**4**
1.1. Was easy to implement		8	18	5
1.2. took more time than I expected	6	9	5	11
1.3. is something I could easily employ in my teaching approach		6	20	4
1.4. is something I do employ in my daily teaching practice	5	16	5	4
1.5. is consistent with the typical teaching practice in my country	8	6	4	
1.6. could potentially serve summative assessment purposes	6	6	10	8
1.7. could produce information that may inform my instructional decisions about my next steps in the unit		2	21	9
Students were:				
1.8. used to the peer/self- assessment method	10	8	10	2
1.9. keen to engage in the peer/self- assessment method	1	5	16	8
1.10. able to reflect on the merits of the artefact they were assessing	4	13	11	2
1.11. able to improve their artefact as a result of the feedback comments		8	18	4

This paragraph focuses on the synthesis of the results of what the teachers considered as the main strengths and weaknesses of the assessment method of peer assessment. Teachers stated that this method provides opportunity to the students to reflect on their initial artefact and to compare their learning with that of their classmates. Moreover, the teachers mentioned that some students considered the feedback comments and used them in order

to improve their answer. Furthermore, all the students can benefit from this method because they can express their opinion. Additionally, the assessment criteria became clear to all students. Another advantage of peer assessment reported by the teachers, is that it can help them gain insights into their students' misconceptions and difficulties. Hence, this complements the teacher's role in assessment. One of the teachers claimed that formative assessment could serve summative purposes. The following quote illustrates this:

> *Formative assessment may precede summative (formative comment on first, then summative mark). The learning task/problem task which was assessed formatively could be used again with some time delay and assessed summatively.*
>
> (Czech Republic)

The challenges emerging from the data refer to the need to:

a) make peer assessment less time-consuming

> *Peer-assessment requires more time than would be used in the normal school for the same subject.*
>
> (Switzerland)

b) help students develop the competence of assessing their classmates' work,

> *Students are given the opportunity to rethink about their work again. You get a particular feedback on whether their representation of the facts is understandable and readable*
>
> (Switzerland)

> *Yes is useful for Physics and perhaps other subjects as well. It was fun. You also have the chance to correct your work based on the peer-feedback, before submitting it to the teacher. Also it was good to give feedback to others and learn from their mistakes*
>
> (Switzerland)

c) train students to be reliable while giving feedback comments without being influenced by the social context that is attached to each pair of students,

> *I would prepare my students for this assessment method at first, give them opportunity to discuss problematics of assessment, speak with them about their feelings and whether they understand the assessment…*
>
> (Czech Republic)

d) manage the time being dedicated to the process of peer feedback.

> *Also we had time limitations to proceed to further revisions*
>
> (Switzerland)

The quote below is indicative of the second challenge: "Sometimes students are not able to find the mistakes." (Czech Republic). The following quote exemplifies the third weakness:

> *Since the assessments were quite clear, the students did not develop any intrinsic motivation for the assessment, but just did it as part of the project*
>
> (Switzerland)

> *Lack of factual knowledge or interest in learning make the peer-assessment non-effective, sometimes it could be used by students to get even with their peers*
>
> *(Czech Republic)*

One suggestion that teachers provided in order to overcome these challenges includes the development of a specially designed tool to support students on how to give comments to their classmates. The quote below illustrates this:

> *We need specific tools (like template) that help the teacher and students tap into the power of this approach.*
>
> (Cyprus)

> *For practice of the assessment methods I would like to have something like a "dictionary" of assessing terms, so every student in the class will be able to understand the basic terms and students with limited vocabulary will be able to express their assessment more exactly (not only with „black-and-white" point of view).*
>
> *(Czech Republic)*

6.3 On the fly assessment method

Table 66

Results from teacher's self-reporting tool

1. The assessment method "Interactions on the Fly":	**Total**			
	1	**2**	**3**	**4**
1.1 Was easy to implement	1	4	12	2
1.2. took more time than I expected	5	7	5	3
1.3. was beneficial to the students		1	8	10
1.4. was beneficial to myself as the teacher		1	8	10
1.5. could produce information that may inform my instructional decisions about your next steps in the unit		3	8	8
1.6. is something I could easily employ in my teaching approach		2	11	6

1.7. is something I do employ in my daily teaching practice	2	2	6	9
1.8. is consistent with the typical teaching practice of teachers in my country	2	5	6	4
Students were:				
1.9. used to participate in assessment dialogue with the teacher	1	7	10	1
1.10 keen to contribute to the discussions	1	7	9	2
1.11. in a position to act (in real time) on the feedback they received during the discussion in real time		5	12	2

The teachers' responses suggest that most of the teachers felt that it would be possible to incorporate this method into their teaching practice and identified strengths and opportunities, as illustrated in the following indicative quotes:

> *Gives real time insights into how students are doing and facilitates very productive feedback.*
>
> (United Kingdom)

> *It was possible to provide instant feedback that could enable the teacher to move learning in a particular direction.*
>
> (United Kingdom)

> *Depending on the students' needs, it is possible to change the focus of the discussion to either narrow it down to a particular area or open it up to take a broader view.*
>
> (United Kingdom)

> *Allows much more 'space' for assessment during the lesson, enabling the teacher to find out things in discussion that would never appear in a final product that typically emerges in formal assessment.*
>
> (United Kingdom)

> *Continual feeding of information to inform on the progress and teaching and learning process. This gives the teachers the opportunity for differentiating his or her approach, either during the current lesson or for subsequent lessons.*
>
> (Cyprus)

Speaking to students in a more informal way and asking questions directly about what they are doing. This is not always possible in a normal class lesson. In this way, it allows assessment without pupils knowing one is assessing them, also on skills not covered in written tests.

> *It allows the right kind of pedagogy which starts from the student's own knowledge.*
>
> (Finland)

> *It gives quick feedback to the student. The teacher becomes aware of the student's knowledge and skills and possible problems. The method engages students.*
>
> (Finland)

> *An evaluation that is based on the students' reasoning. It lets me adjust the evolution of the teaching sequence according to the improvements of the students. Students are actors, it's them who construct, propose.*
>
> (France)

> *It does allow the students to make explicit to others their hypothesis and to get more involved [in the task]. They can discover too the weak points of their argumentation. This allows the teacher to better know and understand where the students are.*
>
> (France)

However, the teachers identified some weaknesses of the assessment method along with challenges they met when implementing it. Next, we provide relevant, indicative quotes:

> *The teacher needs to be focused on what s/he is assessing and not get distracted by the task.*
>
> (Cyprus)

> *It is not likely to be able to discuss with all students in the class and assess them within a single lesson.*
>
> (United Kingdom)

> *In cases of students working in pairs or groups, it is difficult to assess them properly, since the teacher will have to move around in the room and will not be able to spend sufficient time with each group to collect adequate assessment information about different aspects of the task at hand.*
>
> (United Kingdom)

There is an issue in relation to the validity of the judgement and how this can be recorded. Also, there is an issue as to whether students need to know about what judgement is being made for them.

(United Kingdom)

This assessment method is time consuming. There is already an issue with time since the curriculum is very loaded; providing such feedback would require even more time.

(United Kingdom)

It demands time and energy. There should be more co-operation with the teachers, if not, students are confused of different assessment methods.

(Finland)

Is this method reliable? The successfulness depends on students' activity and willingness to take part in discussions.

(Finland)

"[Challenges] to organize and lead the classroom discussion with the students while trying to guide them to a specific goal without they feel so, I mean by that to give the impression that the direction taken by the discussion comes from the students"

(France)

It takes time

It's live and we don't always have the vocabulary or the adapted argumentation

(France)

Difficulties to make a personal follow-up

(France)

In addition, some teachers reported challenges they encountered while enacting this assessment method. For example, they found it hard to address the diversity of levels of ability within the class: *"There is a variety of levels of ability within students and they also hold many different ideas about the topic. It is challenging to handle them and respond accordingly"* (United Kingdom). Also, they found assessing students' reasoning harder than other competences in particular contexts: *"the hardest thing to assess is the actual reasoning and the reasoning in this particular case is actually really complicated, in my*

opinion" (United Kingdom). Further, an upper secondary teacher stated that *"if a student group is big it is hard to get such information from individual students which can be used as a reliable basis for assessment. That's true especially if the knowing of students is not good." (Finland)*

Another teacher admits that having a path in mind herself persuaded her to ask questions that were too leading or giving leading feedback: *"I found myself saying a lot more things, like, yes, that's right, yes, that's good, rather than I think in the last time … because this time I had one track that was really clear to me, I think it probably felt a lot more to some of them like there's a right and wrong"* (United Kingdom).

Another challenge relates to the need to support students in real time, which is a clearly demanding task: "*You have to either really know the problem well, you know, have done it lots and really understand all the different approaches and different mistakes and skills that you need and stuff, or be really quick on your feet with just being able to look at their weird sort of calculations and say, oh, I can see where that came from and I think teachers are really scared of that kind of thing."* (United Kingdom). Another teacher stated that *"You need to have very good background on the subject in order to be able to understand what the students are saying (meaning) and at the same time to be able to think how to respond in a way that could help them move forward"* (Cyprus). Another teacher wrote that it can be challenging to *"find correct words and questions in a situation and to ask questions which are open enough"* (Finland) Further, the importance of teachers' professional development and experience was emphasised: *"Teachers should have enough experience as a teacher, good content knowledge and good knowledge of the curriculum to successfully implement the method"* (Finland).

The teachers also suggested specific resources that could be useful for supporting the uptake of this assessment method, as follows:

- Generic guidelines (e.g. probing questions) for how to effectively promote inquiry dialogue.
- More opportunities to practice the assessment method.
- Examples and ideas as to how other teachers have implemented this assessment method.
- Resources to record data or monitor students' progress
- A clear assessment framework for each task so that data can be easily populated into a template

6.4 Structured assessment dialogue

Table 66

Results from teacher's self-reporting tool

1. The assessment method "Structured assessment dialogue":	**Total**			
	1	**2**	**3**	**4**

1.1 was easy to implement		1	3	1
1.2 took more time than I expected	1	3	1	
1.3 was beneficial to the students		1	2	2
1.4 was beneficial to myself as the teacher		1	3	1
1.5 is something I could easily employ in my teaching approach	1	2	2	
1.6 is something I do employ in my daily teaching practice			2	3
1.7 is consistent with the typical teaching practice of teachers in my country			3	2
2. The assessment tool I was provided with and used during the enactment of the structured classroom dialogue:				
2.1 consisted of clearly formulated items		1	3	
2.2 contained items expressed in an appropriate format (e.g. likert-scale).		1	3	
2.3 enabled the collection of data that provided reliable indications about the students' level of attainment of the targeted competence		2	1	
2.4 included unnecessary statements/questions.		4		
2.5 failed to capture important assessment information.		3	1	
3. During the structured classroom dialogue, the students were:				
3.1 keen to engage in the process		1	2	2
3.2 enacted effectively the roles they were assigned.			4	1
3.3 followed effectively the relevant rules.			4	1

This paragraph focuses on the synthesis of the results of what the teachers considered as the main strengths and weaknesses of the assessment method of structured assessment dialogue. The teachers mentioned that this method engages the whole class and helps students learn from their classmates. Also, they stated that the method is characterised by clear rules, short time of execution. In addition, they valued this assessment method highly in terms of its facility to serve as a useful formative assessment means. Moreover, they referred to the fact that this method provides teachers with the opportunity to easily and quickly identify where the students are. The quote below is illustrative:

> *It is possible to get a nuanced picture of the students' understanding – something that a written text would not be able to capture to the same extent.*

(Denmark)

Regarding the weaknesses, teachers mentioned that the effectiveness of this assessment method seems to depend, to a certain extent, on the level of the engagement of the class

and the preparedness and willingness of the students to genuinely participate in the whole process. Another point that was raised relates to the high demands placed on the teachers in terms of preparation time but also time needed to actually enact the process in the class. The following quotes indicate this point:

> *It is not something that can be done very frequently as it relies on topics and classroom activities, does need time to set up within a lesson so potentially taking time away from something else.*
>
> (United Kingdom)

> *It can be difficult to make as limited an assessment that it is possible to keep the time frame. It is valuable as an alternative assessment method / teaching method.*
>
> (Denmark)

Another weakness that was recognized by the teachers is the following:

> *To keep the drive-over-time in the method so that the feedback group is serious about their own learning (self-evaluation). To choose the "appropriate" students to take a seat in "the hot chair". To make the demands of the students and their roles easy-to-read for everybody.*
>
> (Denmark)

Some teachers noticed that the students did not find this method useful. The following quote exemplify this limitation:

> *There is a tendency that students may experience it as an interrogation. In that case, the students will not see it as being useful with a view to the future.*
>
> (Denmark)

One suggestion that teachers provided in order to overcome these challenges includes the training and frameworks to work. The following quote illustrates this:

> *They need some pre-described units that one could get to try out directly, in which there are also ideas for the questions.*
>
> (Denmark)

If the teachers had the opportunity to repeat the implementation of this assessment method, they mentioned that they would make certain changes to specific parts. In particular, they indicated that they would invest more time in thinking through the roles that are to be assigned to the students and that they would assume more flexibility with respect to the time limit that would be given to the students to perform the various parts of the process. The following quotes are indicative of these suggestions.

> *…Think more carefully about which students took which roles and how it was explained to students to try and ensure they could focus on the task more clearly.*

(United Kingdom)

Allocating different roles to the students in the feedback group, so that each student gets his/her own assignment such as "focus on definitions, units and sizes" or "focus on the use of specific content knowledge".

(Denmark)

I probably would not obey the five minutes, but use the time that is needed on the dialogue.

(Denmark)

6.5 Synopsis

In general, teachers recognize that when implementing one of these formative assessment methods, they have the opportunity to collect data which could inform them about students' performance in inquiry competencies and also highlights their students' difficulties and needs. Interpreting these data allowed teachers to adapt their instruction in order to correspond accordingly. In those methods that occur during the course of an inquiry this also allows students to question the ways they are working and to change and improve their approach. It needs to be noted here that each of the four assessment methods added to the classroom assessments that teachers were already doing and so supplemented their understanding of student performance in inquiry competencies. That is to say that these four assessment methods are not to be regarded as stand-alone methods but rather form a repertoire of methods that teachers can draw from to use at chosen points in the student learning. This is necessary both to inform summative judgements about student performance of inquiry competence across a range of contexts and particularly to help teachers provide formative guidance.

One factor that teachers raised about implementing formative assessment methods into their teaching practice is that it is demanding time-wise, especially for these assessment methods that are applied to individual students. One suggestion that teachers provided in order to overcome these challenges includes the formulation of guidelines or tools that could potentially support them on how to apply each assessment method more effectively.

7. Synthesis of key findings from classroom trials

In this part we elaborate on certain tentative claims that seem to emerge from the data that have been presented in the previous sections of the report. These claims seek to capture key ideas that could be brought to bear on the effective enactment of the assessment methods that have been tested in WP5. Our intent is to shift the discussion to a level of generality that could transcend specific enactments of individual assessment methods.

1. We found that teachers and students benefitted when the focus of classroom assessment centred on formative feedback. Teachers could provide feedback to their students using scaffolds and tools to achieve this.

The data processed in WP5 suggest that it is possible to develop tools/scaffolds that could help teachers provide feedback that attends to the important aspects of the targeted learning objectives in a systematic manner. For instance, in the case of written feedback the teachers used tools in order to facilitate the enactment of the respective assessment method, which seemed to have helped them provide feedback that addressed the relevant, pre-specified criteria. The tool also helped the teachers to systematize the evaluation of the students' artefacts by diagnosing students' difficulties in relation to the specific aspects of the competence of interest in each case. Therefore, it made it easier for them to provide feedback that corresponded to the individual needs of the students.

Even though the focus of the implementations was placed on the teacher rather than on the students, the available data suffice to highlight the importance of this (as reported by the teachers in the data from the self-reporting tools) but also the challenges that seem to emerge in this respect (e.g., as evidenced by the data from peer-feedback). Providing detailed guidance enables students to improve their work. Even the most concise and thoughtful feedback will not be useful if students do not actually act on it. This part of formative assessment is at least as important as the provision of useful feedback (Higgins et al., 2002).

Students can be usefully supported (e.g. by tools involving reflective probes), so as to engage with the process of (a) interpreting the feedback comments they receive and (b) deciphering specific directions for immediately next steps. This increases the likelihood for students to act on the feedback comments they receive in a thoughtful and productive manner.

2. The tools developed by LWGs across the four assessment methods enabled teachers to make assessment criteria specific and explicit and this both supported teachers in providing feedback and students in responding to guidance.

The assessment of inquiry competencies is relatively new in most countries and the assessment tools produced supported teachers in defining and articulating what a quality performance consisted of. These tools provided teachers with both support for making judgments and also the means for devising suitable and appropriate feedback comments for students, thereby strengthening the assessment literacy of both parties. The tools included

descriptions of performance for specific inquiry activities, rubrics, checklists and learning progressions.

3.**Teachers need time and support in evolving and developing these formative tools to fit their classroom contexts**

It is not feasible to anticipate the diversity of possible inputs by students and prepare corresponding guidance notes for concrete next steps. Even if this was indeed possible it would still be unproductive from the perspective of the teaching/learning process, for two reasons: (a) it runs the risk of reducing the task of formulating feedback comments to an algorithmic act, totally neglecting the need to engage teachers in a creative manner; (b) it does not address the need to customize feedback (both its substantive and affective components) by tailoring it to the needs of individual students. Furthermore, this approach provides teachers with the opportunity to reflect on the use and purpose of the tool and provides a sense of ownership of the assessment process.

4. Teachers need to establish and sustain a classroom culture in which formative exchanges can function

The extent to which any assessment method can serve as a valuable teaching and learning experience (both for teachers and students) is contingent on the presence (or lack thereof) of an appropriate "assessment culture" within the class. This has been demonstrated by the data presented in this report. A piece of evidence that could serve to illustrate this, relates to the assessment method that drew on peer-feedback. The available data seem to be suggesting that helping students frame the whole process in a manner that they could perceive to be valuable or useful, is of paramount importance for facilitating productive enactment of the two roles (i.e., peer-assessor and peer-assessee). For instance, in some cases the data reported in WP5 suggested that students might end up enacting the role of the assessor in a non-thorough manner, without taking the effort to reflect on the strengths and weaknesses of their peers' work in a critical and rigorous manner. In a similar manner, our data provide evidence suggesting that when working as peer-assessor they might dismiss useful comments provided by their peers, or avoid taking the challenge posed by their peers in terms of justifying or reconsidering aspects of their work. At the same time, it is important to highlight that there is evidence in our data suggesting that providing students with sufficient opportunities to exercise this assessment method could foster a more productive stance on the part of the students which, in turn, could enhance the learning benefits associated with this assessment method.

5. There is a need for professional development programmes for teachers to become better positioned to cope with the intricacies, challenges and complexities involved in implementing formative assessment

The learning environment is a tremendously rich source of assessment data. The teacher is continuously confronted with various inputs pertinent to the students' current state of learning. One challenge inherent in any attempt to use these inputs formatively, relates to the need to diagnose needs and difficulties evident in these often not well-articulated inputs. Another challenge includes making decisions about what would be appropriate

feedback in particular situations. This could become even more pressing in cases when these decisions need to be made in real time (e.g. when using interactions on the fly as a means of formative assessment). Managing this complexity is certainly not an easy task. It requires the accumulation of relevant experiences but also practices and skills, which could supply teachers with theoretically- and experientially-informed lens for analysing and responding to the feedback they are exposed to in the learning environment. This requires regular opportunities for teachers to exchange ideas and share practice with peers guided and supported by teacher educators and researchers.

References

Araceli Ruiz-Primo, M., & Furtak, E. M. (2006). Informal formative assessment and scientific inquiry: Exploring teachers' practices and student learning. Educational Assessment, 11(3-4), 237-263.

Gielen, S., Peeters, E., Dochy, F., Onghena, P., & Struyven, K. (2010). Improving the effectiveness of peer feedback for learning. Learning and Instruction, 20(4), 304-

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81-112.doi:10.3102/003465430298487

Higgins, R., Hartley, P. & Skelton, A. (2002). The conscientious consumer reconsidering the role of assessment feedback in student learning, Studies in Higher Education, 27(1), 53–64. doi:10.1080/03075070120099368

Orsmond, P., Merry, S., & Reiling, K. (2000). The use of student derived marking criteria in peer and self-assessment. Assessment & Evaluation in Higher Education, 25(1), 23-38

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, *18*(2), 119-144. doi:10.1007/BF00117714

Tsivitanidou, O. E., Zacharia, Z. C., & Hovardas, T. (2011). Investigating secondary school students' unmediated peer assessment skills. Learning and Instruction, 21(4), 506-519.

Värlander, S. (2008). The role of students' emotions in formal feedback situations. *Teaching in Higher Education*, *13*(2), 145-156. doi:10.1080/1356251080192319

Appendix

Appendix I: Written feedback

Examples of tools of providing feedback by the teacher

Germany

Feedback for [student name]

This is the goal! Your experimental plan should include the following aspects:	This is what you have already achieved! In your experimental plan, you have considered the following aspects:				This is what you should consider! In your next experimental plan, you should consider the following aspects:
	Not at all	Partly	Completely	Justification	
The experiment allows for completely testing the hypothesis.	☐	☐	☐		
The dependent variable is correctly specified ...	☐	☐	☐		
... and is observed using an appropriate measuring tool.	☐	☐	☐		
The independent variable is correctly specified, ...	☐	☐	☐		
is described in detail using indications of size or amount ...	☐	☐	☐		
and is varied under controlled conditions.	☐	☐	☐		
Confounding variables are taken into account in order to keep the experimental conditions constant.	☐	☐	☐		
A control experiment is carried out to confirm the results.	☐	☐	☐		
Your list of materials and chemicals is complete.	☐	☐	☐		
You have considered all necessary safety precautions.	☐	☐	☐		

Denmark

This is the learning objective You ought to show that you have achieved the following learning objective:	**This is what you have achieved** In your assignment				**This is what you ought to consider** In your next assignment
	To a very small extent	To some extent	To a large extent	Explanation	

Tool for Coding Teachers' Written Feedback Comments

A. Metadata
1. For each separate sample group (e.g. a class or a teacher), summarise the instructions given by the teacher for the students' artefacts. 2. Include the feedback form used by the teacher or explain how you prompted the teacher to give feedback 3. Were the students at some point presented with learning objectives and/or learning progressions? YES/NO
4. What are the competence goals (learning objectives) and how do they relate to the ASSIST-ME competence(s) of interest 5. Write a checklist of relevant aspects of the competence goals (could be part of A.4)
6. What are the levels of attainment (or learning progression steps) that relate to the learning objectives? (could be part of A.4)

B. The teacher's assessment of the student's level of attainment				
1. Given the "checklist" of aspects in A.5, to what extent did the feedback comments take into account these aspects				
The feedback comments did not take into account any of the relevant aspects 1	The feedback comments took into account only some of the relevant aspects 2	The feedback took into account half of the relevant aspects 3	The feedback comments took into account most of the relevant aspects 4	The feedback comments took into account all relevant aspects 5
Briefly justify why you have assigned this score on the given Likert scale:				

2. Which level of attainment did the teacher give to the student **(OPTIONAL)** [This assumes that the teacher is asked to assign a level of attainment (from those specified in the corresponding teaching scheme of the Local Working Group) to the student, but also to provide a relevant justification. This information does not need to be communicated to the students.]	
3. Group consensus (at the level of the LWG) about the level of attainment of the student **(OPTIONAL)**	

C. Justification offered by the teacher to the students about their level of attainment					
1. Were the learning objectives part of the feedback? (e.g. as a part of a standard feedback form) YES/NO					
2. Specify the total number of statements within the feedback comments that contain references to either (a) what the student(s) has(have) achieved or has(have) yet to achieve (according to the levels of attainment). In each case, specify whether this statement is justified by the teacher **(OPTIONAL)**					
References to what the student **<u>has already achieved</u>** with respect to the targeted competence			References to what the student **<u>has yet to achieve</u>** with respect to the targeted competence		
Number of references that are justified by the teacher (i.e. the teacher gives a reason for it)	2a	Illustrative quote:	Number of references that are justified by the teacher (i.e. the teacher gives a reason for it)	2c	Illustrative quote:
Number of references that were not justified by the teacher (i.e. the teacher did not give a reason for it)	2b	Illustrative quote:	Number of references that were not justified by the teacher (i.e. the teacher did not give a reason for it)	2d	Illustrative quote:

3. To what extent did the teacher justify his/her references to what the student **has already achieved** with respect to the targeted competence?

Not applicable: it did not include any such references	References were all without justification	References were mostly without justification	Balanced	References were mostly justified	All references were justified
0	1	2	3	4	5

Briefly justify why you have assigned this score on the given Likert scale:

4. To what extent did the teacher justify his/her references to what the student **has not yet achieved** with respect to the targeted competence?

Not applicable: it did not include any such references	References were all without justification	References were mostly without justification	Balanced	References were mostly justified	All references were justified
0	1	2	3	4	5

Briefly justify why you have assigned this score on the given Likert scale:

D. Guidance provided by the teacher to the student(s) on how to proceed				
1. Does the student need guidance (in reference to the "checklist" of the relevant aspects)?	YES/NO			
2. Was the guidance indeed related to the competence of interest? Yes/No				
3. To what extent does the feedback provide specific guidance to student about concrete next steps?	To no extent (i.e. non-specific guidance)			To a high extent (i.e. very specific guidance)
	1	2	3	4
Briefly justify why you have assigned this score on the given Likert scale:				

E. Affective dimension To what extent does the feedback carry connotations relating to affective aspects?		
1. Specify the frequency of statements (e.g. sentences or segments of sentences) within the feedback comments that carry affective connotations (explicated through specific statements/terms such as bravo, well done etc). **(OPTIONAL)**	Frequency	Illustrative quote
(i) Frequency of these instances that were likely to serve as encouraging feedback but were restricted to vague phrases (e.g. excellent/bravo etc.)	1a	
(ii) Frequency of these instances that were likely to serve as encouraging feedback but also identified (and provided credit for) what the student has achieved	1b	
(iii) Frequency of these instances that were likely to serve as discouraging feedback and were formulated in a vague manner (e.g. you have not put much effort, this is substandard, this was not a good response etc.)	1c	
(iv) Frequency of these instances that were likely to serve as discouraging feedback and were also formulated in a specific manner that identifies what the student has failed to do.	1d	

2. Were there few or many affective comments in the feedback?

None or few affective comments			Many affective comments
1	2	3	4

3. Was the overall feedback positive or negative

Mostly negative		Balanced		Mostly positive	Does not apply
1	2	3	4	5	0

F. Complexity of the language in the feedback

1. How difficult was the language in the feedback (in terms of sentence structure and/or terminology)?

Not difficult			Very difficult
1	2	3	4

Briefly justify why you have assigned this score on the given Likert scale:

G. The students' usage of the feedback <u>in the second artefact</u>

1. To what extent did the student address the feedback (on the first artefact) in his/her second artefact?

The student did not act on any aspects in the feedback comments, which could have been applied in the second product.			The student acted on all aspects in the feedback comments, which could have been applied in the second product.
1	2	3	4

Briefly justify why you have assigned this score on the given Likert scale:

Teacher-reporting tool

This tool is divided in two parts. The first includes a total of 12 items associated with the assessment method "Marking (Grading & Written Comments)". The items are expressed in a likert scale and are grouped in three categories depending on whether they refer to the perceptions of the teacher about (a) the process of enacting the assessment method, (b) the features of the assessment method more broadly and (c) how the students engaged with the assessment method. The second part includes four open-ended questions.

Name: **Date:**

School:

Part A

Use the 4-point scale shown below to indicate the extent to which you agree or disagree with each statement in the table.

1	*2*	*3*	*4*
Not at all			*To a great extent*

1.The assessment method "Grading & Written Comments":				
1.1. was easy to implement	1	2	3	4
1.2. was easy to interpret students' difficulties and needs from the artefacts they created	1	2	3	4
1.3. was easy to write descriptive comments that could be helpful to the students	1	2	3	4
1.4. took more time than I expected	1	2	3	4
1.5. was beneficial to the students	1	2	3	4
1.6. was beneficial to myself as the teacher	1	2	3	4

1.7. is consistent with my usual teaching approach in my class	1 2 3 4
1.8. is consistent with the typical teaching practice in my country	1 2 3 4
1.9. could potentially serve summative assessment purposes	1 2 3 4
1.10. could produce information that may inform your instructional decisions about your next steps in the unit	1 2 3 4
2. The students were:	
2.1. used to work with individualized descriptive feedback comments	1 2 3 4
2.2. in a position to interpret the feedback they were provided with	1 2 3 4
2.3. kkeen to engage with reflection on the comments they received from their teacher	1 2 3 4
2.4. keen to take into account the feedback comments in revising their artefacts	1 2 3 4

Part B

Please respond to the following four questions.

What are the main strengths and weaknesses of Marking (Grading & Written Comments) as an assessment method?
What challenges and opportunities do you identify in Marking (Grading & Written Comments) as a formative assessment method?
Opportunities: Challenges:
What resources could help teachers overcome these challenges?
What would you have done differently, in terms of structure and procedures you have employed, if you had the opportunity to repeat the implementation of this assessment method?

Tool with illustrative examples of written feedback comments

The experimental design includes	Coding Category	Feedback Notes

Identification of variables related to the phenomenon	All variables that students are expected to identify are mentioned	*Good work, you have identified enough variables*
	Some of the variables that students are expected to identify are mentioned	*Good that you have identified some variables. However, there might be some more variables influencing the phenomenon that you need to consider.*
	No variables that students are expected to identify are mentioned	*You need to try more and think about variables that when altered will influence the phenomenon.*
	Other processes are mentioned as variables (eg. "The number of measurements I will take")	*Do you think that this is a variable of the phenomenon itself or something related to the accuracy of our experimental method for taking measurements?*
	Discrimination between simple (i.e. mass) and complex variables (i.e. density)	*It's good that you have identified enough variables that might influence the phenomenon. Additionally, think of the possibility that not all of them are independent but some might covariate.*
Identification of the independent variable	The independent variable is defined	*Good that you have identified which is the variable that you are going to change.*
	No independent variable is defined	*You need to think which is the variable that you are going to change, in order to be able to answer, whether it is influencing the phenomenon or not.*
	More than one independent variables are defined	*You have identified more than one variable that you are going to change. Doing so, how are you going to be sure, which of these influenced the phenomenon?*
	There is confusion between independent and dependent variable	*You need to think again, which is the variable you need to change purposely in order to check whether it influences the phenomenon?*
Valid way of changing the	Valid	*Good, this is a valid way to change the variable A.*

independent variable	The chosen way of changing the independent variable doesn't allow having an accurate measurement of the change	*Good that you have describe a way to change variable A. However, maybe you need to differentiate a bit your method in order to have a way to measure the amount of this change.*
	The chosen way of changing the independent variable inevitably causes alteration on other variables that should remain constant	*Good that you have describe a way to change variable A. However, you need to take care that when changing this variable, no other variables that should remain constant are changing also.*
	There is no description of how the independent variable will be changing	*You need to describe how the variable A will be changing in order to check whether it influences the variable B.*
Identification of the dependent variable	The dependent variable is defined	*Good that you have identified which is the variable that you are going to measure.*
	No dependent variable is defined	*You need to think again, which is the variable that you are going to measure in order to be able to answer whether it is influenced or not when changing the variable A.*
	More than one dependent variables are defined	*You have identified more than one variable that you are going to measure. You need to think again, which is the particular variable that you need to measure, in order to be able to answer whether it is influenced or not when changing the variable A.*
	There is confusion between independent and dependent variable	*You need to think again, which is the particular variable that you need to measure in order to be able to answer whether it is influenced or not, when changing the variable A.*
Valid way of measuring the dependent variable	Valid	*Good, this is a valid way to measure the variable B.*
	The chosen way of measuring the dependent variable	*Good that you are describing a way of measuring. However, you need to think if what you measure is indeed the variable B.*

	doesn't really measure this variable.	
	The chosen way of measuring the dependent variable doesn't allow having an accurate measurement of the change caused	*Good that you are describing a way of measuring variable B. However, maybe you should differentiate your method in order to have more accurate measurements of the change caused on variable B.*
	There is no description of how the dependent variable will be measured	*You need to think of a way for measuring variable B in order to be able to check whether is influenced when changing the variable A.*
Valid way of keeping constant other variables	Valid	*Good that you are describing how you will keep other variables that might influence the phenomenon constant.*
	Not all variables that might influence are kept constant	*Good that you are describing how you will keep other variables that might influence the phenomenon constant. However, look again the table with the identified variables that might influence the phenomenon and check if you have taken care for all of them to be kept constant.*
	The described way of keeping constant other variables doesn't ensure that they are indeed kept constant	*Good that you are describing how you will keep other variables that might influence the phenomenon constant. However, think more carefully whether are indeed kept constant in this way.*
	No care for keeping constant other variables that might influence the phenomenon is taken	*Look again the table with the identified variables that might influence the phenomenon and describe how you will keep these variables constant.*
	Variables already found not to affect the phenomenon are mentioned as not needed to be kept constant	

Appendix II: Self and Peer feedback

Examples of tools for providing feedback by the peer

Denmark

Assessor	
Assessee	
Title of the teaching unit	
Lesson numbers	
Product	
Learning objective	
Progression step 1	
Progression step 2	
Progression step 3	
Progression step 4	
Assessment of progression step	
Justification of the assessment	
What should be kept?	
What should be improved?	
Concrete next step	

Czech Republic

①Did student answer correctly the question: Do plants need the light?

YES			**NO**	

Did the student explain his/her answer?

YES		Why?
NO		

②Did the student explain correctly how necessity of the light is manifested?

YES		Why?
NO		

③Learn the student's image.
Is the experiment designed correctly? Would you like change anything?

YES		I would like to change:
		Why?
NO		

④Aids needed for experiment.

Are aids suggested properly?

YES	
NO	

Are some aids missing or are they useless?

Absent:
Useless:

⑦Is it possible to use this experiment to verify if plant needs light?

YES	
NO	

Why?

⑧Did the student write correctly other factors appropriate for the plant growing?

YES			**NO**	

⑨Which mark will you assess the whole protocol with?

⑩Justify your choice:

Assessment of design of experiment

The assessment should contain list of mistakes not their solution!

Are factors in table selected properly?

YES	*Advice to your peer if the selected factors are correct or s/he forgot some of them:*
NO	

Did the student select the factor appropriate for the performing of experiment?

YES		*If NO, advise, why the selected factor is not correct:*
NO	⇨	

Is the impact of this factor adequately explained?

YES		*If NO, what is needed to add or explain better?*
NO	⇨	

Is the experiment designed properly?

YES		*If NO, write where is the mistake:*
NO	⇨	

Should the design of experiment be modified to be right?

YES	⇨	*If YES, advise what should be improved:*
NO		

Are aids suggested properly? (If NO, write into the box which aids should be added or removed.)

YES		- *add:* - *remove:*
NO	⇨	

Which mark will you assess the whole protocol with?

	1	**2**	**3**	**4**	**5**
Mark with cross:					

Justify your choice:

Cyprus

Model Assessment	
1. Name of model:	
2. Modeler's group name:	
3. Date:	
4. Phenomenon represented by your model:	
5. Criterion 1,2,3,4 : representation of phenomenon **To what extent does the model incorporates all necessary components associated with the operation of the target phenomenon (objects, variables, processes and relationships)? Specifically identify missing components.**	
6. Criterion 5: Explanatory power of the model **To what extent does your peers' model includes a mechanism that can help one account for operation of the target phenomenon? Justify your response and if you think that it does include such a mechanism offer a brief description of the mechanism.**	
7. Criterion 6: Predictive Power **To what extent can your peers' model be used by someone to predict the operation of the phenomenon under specified (not previously observed) conditions? If yes, formulate a prediction and justify how it can be derived from your peers' model**	
8. How would you suggest your peers to revise/change their model so as to make it more powerful in terms of the six criteria?	

TOOL FOR CODING STUDENTS' PEER FEEDBACK COMMENTS

(To be used with Research Design V, which focuses on the assessment method *Self & Peer-feedback*)

A. METADATA
1. For each separate sample group (e.g. group of students), briefly characterize the artefact that was assessed by the student/s. Which prompts were used to elicit feedback from the assessor in the tool?
2. Choose one of the following options, which apply in your case. The peer-feedback comments were: ☐ Structured (usage of rubrics with specific assessment criteria) ☐ Unstructured (no rubrics were given to the assessors)
3. Name and describe the competences that were under emphasis in this assessment.
4. Provide the list of criteria or learning objectives that students were asked to attend to while assessing their peer's artefact. Offer a brief description of each criterion.

5. The peer feedback comments were exchanged between:	individual students		groups of students	
			number of students in each group	

B. AFFECTIVE DIMENSION To what extent does the feedback carry connotations relating to affective aspects? (Note: B1 is **optional** whereas B2 is **mandatory**)		
1. Specify the frequency of statements (e.g. sentences or segments of sentences) within the feedback comments that carry affective connotations (explicated through specific statements/terms such as bravo, well done etc.). **(OPTIONAL)**	Frequency	Illustrative quote
(a) Frequency of these instances that were likely to serve as encouraging feedback but were restricted to vague phrases (e.g. excellent/bravo etc.)		
(b) Frequency of these instances that were likely to serve as encouraging feedback by explicitly identifying (and providing credit for) what the student has achieved		
(c) Frequency of these instances that were likely to serve as discouraging feedback and were formulated in a vague manner (e.g. you have not put much effort, this is substandard, this was not a good response etc.)		
(d) Frequency of these instances that were likely to serve as discouraging feedback and were formulated in a specific manner that identifies what the student has failed to do.		

2. Overall, what is the likely affective impact of the feedback comments on the student taking into consideration other data sources e.g. post-instructional interviews, observations? Select the level of the Likert-scale you deem more appropriate. **(MANDATORY)**

Likely to have a very negative impact	Likely to have a rather negative impact	Neutral (not likely to have any affective impact)	Likely to have a rather positive impact	Likely to have a very positive impact
1	2	3	4	5

Briefly justify why or how you have selected to assign this score:

C. COMPREHENSIVENESS AND VALIDITY OF FEEDBACK COMMENTS

1. To what extent did the feedback comments take into account the intended criteria (specified in A.2)? (select the appropriate level of the Likert scale)

The feedback comments did not take into account any of the criteria	The feedback comments took into account only some of the criteria	The feedback comments took into account most of the criteria	The feedback comments took into account all the criteria
1	2	3	4

If the assessor has only taken into account some of the criteria (which means that you have selected either 2 or 3 in the likert scale) specify which those criteria were:

If you selected level "1" on the Likert scale then only respond to question 2 and proceed to part D. Otherwise, respond to questions 3 and 4 and then proceed to part D.

2. Where did the assessor place the emphasis instead? Select one or more of the options below (each option denotes an alternative criterion to which students are likely to pay attention to):

(a) superficial aspects associated with the appearance of the artefact (e.g. size of diagrams, length of text)	

(b) the level of complexity of information provided by the artefact (e.g. tendency to privilege technical terms)	
(c) the level of detail provided by the artefact	
(d) the quality of textual information in terms of grammar/syntax/spelling.	
(e) other *specify:*	

This could be completed separately for each criterion (out of those specified in A2), **if this applies.**

You might need to create additional copies of table C (Q3-4) as needed.

3. To what extent did the peer-assessor draw on the criteria/criterion in a thorough manner? (select the appropriate level of the Likert scale)

The assessor drew on the criteria/criterion in a very superficial manner	The assessor drew on the criteria/criterion in a rather superficial manner	The assessor drew on the criteria/criterion in a rather thorough manner	The assessor drew on the criteria/criterion in a very thorough manner
1	2	3	4

Briefly justify why or how you have assigned this score on the given Likert scale:

4. To what extent was the assessor's judgment with respect to the criteria s/he attended to, valid? (select the appropriate level of the Likert scale) **(OPTIONAL)**

The feedback comments were totally non-valid	The feedback comments were mostly non-valid	The feedback comments were mostly valid	The feedback comments were totally valid
1	2	3	4

Briefly justify why or how you have assigned this score on the given Likert scale:

D. RELIANCE OF THE ASSESSOR ON THE CRITERIA WHILE CONSTRUCTING HIS/HER OWN ARTEFACT (OPTIONAL)

1. To what extent did the peer-assessor draw on the intended criteria while constructing his/her own artefact (specified in A2)? (select the appropriate level of the Likert scale)

The assessor did not draw on any of the criteria	The assessor drew on some of the criteria	The assessor drew on most of the criteria	The assessor drew on all criteria
1	2	3	4

If you have selected either 2 or 3, specify the criteria that the assessor sought to address while preparing his/her own artefact.

This could be completed separately for each criterion (out of those specified in A2), **if this applies.**

You might need to create additional copies of table D as needed.

If you have selected either 2, 3 or 4 then respond to the following two questions. Otherwise move to part E

2. To what extent did the peer-assessor draw on the criteria in a thorough manner? (select the appropriate level of the Likert scale)

The assessor drew on the criteria in a totally superficial manner 1	The assessor drew on the criteria in a rather superficial manner 2	The assessor drew on the criteria in a rather thorough manner 3	The assessor drew on the criteria in a very thorough manner 4

Briefly justify why or how you have assigned this score on the given Likert scale.

If you have chosen either 1 or 2 in the question 2 (that is to say the feedback comments are totally or rather superficial), then please specify if the feedback comments are sufficient enough for the assessed artifact.

(For example that could be a case in which the assessor provides a positive comment in a specific criterion and it wouldn´t makes sense to elaborate more, because the assessed artifact is sufficient enough in respect to this criterion).

3. To what extent did the assessor address the criteria in a valid manner while constructing his/her own artefact? (select the appropriate level of the Likert scale)

The assessor addressed the criterion(a) in a totally non-valid manner 1	The assessor addressed the criterion(a) in a mostly non-valid manner 2	The assessor addressed the criterion(a) in a mostly valid manner 3	The assessor addressed the criterion(a) in a mostly valid manner 4

Briefly justify why or how you have assigned this score on the given Likert scale.

E. JUSTIFICATION OFFERED BY THE ASSESSOR TO THE ASSESSEE ABOUT THE CRITERIA Please note that E1 and E2 are **optional** whereas sections E3 and E4 are **mandatory**.					
E1. Specify the total number of statements within the feedback comments that contain references to either (a) what the student(s) has(have) achieved or (b) has(have) yet to achieve (according to the criteria). **(OPTIONAL)**					
References to what the assessee **<u>has already achieved</u>** with respect to the targeted competence			References to what the assessee **<u>has yet to achieve</u>** with respect to the targeted competence		
E1a. Total number of such references within the feedback comments.		Illustrative quote:	**E1b.** Total number of such references within the feedback comments.		Illustrative quote:
E2. Specify the total number of statements within the feedback comments that contain references to either (a) what the student(s) has(have) achieved or (b) has(have) yet to achieve (according to the criteria). In each case, specify whether this information is provided in an implicit or explicit manner. (OPTIONAL)					
References to what the assessee **<u>has already achieved</u>** with respect to the targeted competence			References to what the assessee **<u>has yet to achieve</u>** with respect to the targeted competence		
Total number of such references within the feedback comments.			Total number of such references within the feedback comments.		
E2a. Number of references that are justified by the assessor (i.e. the assessor gives a reason for it)		Illustrative quote:	**E2c.** Number of references that are justified by the assessor (i.e. the assessor gives a		Illustrative quote:

			reason for it)		
E2b. Number of references that were not justified by the assessor (i.e. the assessor did not give a reason for it)		Illustrative quote:	**E2d.** Number of references that were not justified by the assessor (i.e. the assessor did not give a reason for it)		Illustrative quote:

E3. To what extent did the assessor justify his/her references to what the student **<u>has already achieved</u>** with respect to the targeted competence? (select the appropriate level of the Likert scale) **(MANDATORY)**

Not applicable: it did not include any such references	References were all without justification	References were mostly without justification	Balanced	References were mostly justified	All references were justified
0	1	2	3	4	5

Briefly justify why or how you have assigned this score on the given Likert scale:

E4. To what extent did the assessor justify his/her references to what the student **<u>has not yet achieved</u>** with respect to the targeted competence? **(MANDATORY)**

Not applicable: it did not include any such references	References were all without justification	References were mostly without justification	Balanced	References were mostly justified	All references were justified
0	1	2	3	4	5

Briefly justify why or how you have assigned this score on the given Likert scale:

F. GUIDANCE PROVIDED BY THE ASSESSOR TO THE ASSESSEE ON HOW TO PROCEED

Guidance could be perceived as helping the assessee move on, which could be either a reference to what the assessee has not yet achieved (implicit guidance) or what the assessee may need to do next (explicit guidance).

This is to be completed separately for each criterion (out of those specified in A2) to which the assessor offered guidance.

You might need to create additional copies of table F as needed.

(Note: Items 1 & 3 are **mandatory,** items 2 & 4-8 are **optional**).

Please specify the criterion: ……

1. To what extent does the feedback provide specific guidance to student about concrete next steps? **(MANDATORY)**	To a very limited or no extent			To a great extent
	1	2	3	4

2. Briefly describe the essence of the guidance that was provided to the student. **(OPTIONAL)**

3. Was the guidance indeed related to the competence of interest? **(MANDATORY)**
Yes/No

If the answer was **no** offer a brief justification.

If the answer was **yes** proceed to answer each of the following four questions as well. Otherwise move to the next instance of guidance included in the feedback comments (if any).

4. To what extent is the guidance valid (in the sense that it could yield significant	Totally non-	Totally valid (it

improvement to the artefact with respect to the criterion under consideration) **(OPTIONAL)**	valid (it will actually undermine the quality of the artefact)			will maximize the quality of the artefact with respect to the criterion of interest)
	1	2	3	4
Briefly justify why or how you have assigned this score on the given Likert scale:				
5. To what extent is the guidance aligned with the current level of the students? **(OPTIONAL)**	Totally non-aligned (either too easy or too difficult extending beyond the reach of the student)			Totally aligned (appropriate balance between guidance and current learning state of the student)
	1	2	3	4
Briefly justify why or how you have assigned this score on the given Likert scale:				

6. To what extent was the guidance formulated in an explicit manner? **(OPTIONAL)**	The guidance was totally implicit 1	2	3	overly explicit/too prescriptive 4
Briefly justify why or how you have assigned this score on the given Likert scale:				
7. To what extent was the level of explicitness appropriate? **(OPTIONAL)**	The guidance was either too implicit or too explicit, providing no or too detailed information to be productive. 1	2	3	The guidance was balanced, providing adequate information to support the students in taking their next steps and to know what is important in doing so 4
Briefly justify why or how you have assigned this score on the given Likert scale:				

8. To what extent was the assessor's reasoning underlying the guidance, revealed to the assessee? **(OPTIONAL)**	Totally hidden			Clearly explicated
	1	2	3	4
Briefly justify why or how you have assigned this score on the given Likert scale:				

Teacher - Reporting tool

This tool is divided in two parts. The first includes a total of 11 items associated with the assessment method "Self & Peer Feedback". The items are expressed in a likert scale and are grouped in three categories depending on whether they refer to your perceptions about (a) the enactment of the assessment method, (b) the features of the assessment method more broadly and (c) students' engagement with the assessment method. The second part includes four open-ended questions.

Name: **Date:**

School:

Part A

Use the 4-point scale shown below to indicate the extent to which you agree or disagree with each statement in the table.

1	*2*	*3*	*4*
Not at all			*To a great extent*

We anticipate that in some cases, you may have only enacted either peer or self feedback. The items include both terms (self and peer). Strikethrough the term (if any) that does not relate to your case.

The peer/self feedback assessment method:				
1.1. was easy to implement	1	2	3	4
1.2. took more time than I expected	1	2	3	4
1.3. is something that I use to include in my daily teaching practice	1	2	3	4
1.4. is consistent with my teaching approach in my class	1	2	3	4
1.5. is consistent with the typical teaching practice in my country	1	2	3	4
1.6. could potentially serve summative assessment purposes	1	2	3	4

1.7. could produce information that may inform your instructional decisions about my next steps in the unit	1	2	3	4
2. The students:				
2.1. were used to the peer/self assessment method	1	2	3	4
2.2. were keen to engage in the peer/self assessment method	1	2	3	4
2.3. were able to reflect on the merits of the artefact they were assessing	1	2	3	4
2.4. were able to improve their artefact as a result of the feedback comments	1	2	3	4

Part B

Please respond to the following four questions

What are the main strengths and weaknesses of self & peer feedback as an assessment method?
What challenges and opportunities do you identify in enacting self & peer feedback as a formative assessment method?
Opportunities: Challenges:
What resources could help teachers overcome these challenges?
What would you have done differently, in terms of the structure and procedures you have employed, if you had the opportunity to repeat the implementation of self & peer feedback?

Appendix III: Interactions on the Fly

Variation within the broader E, S, R, U codes (an optional component of the research design)

As an optional component of the research design for the assessment method "interactions on the fly", some LWGs undertook to further categorize the utterances coded using the ESRU scheme, into subcategories so as to further describe the variation within each of these general categories. The tables below outline a set of subcategories that was proposed and show the corresponding prevalence of each in thee data sets to which it was applied.

expected to use the subcategories shown in the table below. It is important to note that the various LWGs are welcome to offer suggestions for additional subcategories to be appended in this list in cases when they believe that they have encountered utterances that cannot be captured by the existing set of subcategories. For this the LWGs will need to send their suggestion to theWP5 leaders, along with an illustrative excerpt from their data and a brief description for the proposed subcategory.

ELICIT	
E1	Teacher poses a question to elicit students` reasoning about a new (although interrelated) concept/idea/relation
E2	Teacher asks students to offer an example or report data
E3	Teacher repeats the previous question
E4	Teacher asks for clarification
E5	Teacher suggests a false concept/idea/relation and gets students to reflect on
STUDENT	
S1	Student suggests a concept/relation in response to question posed by the teacher
S2	Student offers justification for his/her reasoning
S3	Student provides an example or reports data
S4	Student explicates an inference about an aspect of the topic under discussion

S5	Student poses a question to the teacher related to the topic under discussion
S6	Student provides a “yes/no” answer
S7	Student expresses ignorance
RESPONSE	
R1	Provision of affirmation
R2	Teacher readily offers the right answer to a question posed by him/herself or by a student.
R3	Provision of disconfirmation
R4	The teacher acknowledges a contribution made by the students
USE	
U1	Teacher suggests an activity that could help students resolve a specific (conceptual) issue
U2	Teacher seeks to focus students' attention on something with the intent to facilitate the discussion (e.g. stated opinions/data/examples)
U3	Teacher seeks to engage students in deeper reasoning on something (further analysis/explanation)
U4	Teacher articulates the consensus from series of contributions that were exchanged

Percentages of Elicit subcategories

Instance of dialogue	E1 (%)	E2 (%)	E3 (%)	E4 (%)	E5 (%)
1	69	2	16	11	2
2	57	0	19	24	0
3	60	0	25	15	0
4	54	8	21	17	0
5	75	4	15	6	0

6	74	11	11	4	0

Percentages of Response subcategories

Instance of dialogue	S1 (%)	S2 (%)	S3 (%)	S4 (%)	S5 (%)	S6 (%)	S7 (%)
1	60	10	4	0	0	25	1
2	85	3	0	0	0	2	1
3	63	21	0	1	0	5	0
4	55	5	7	8	3	15	7
5	80	5	1	0	3	11	0
6	37	11	21	13	3	14	1

Percentages of Recognize subcategories

Instance of dialogue	R1 (%)	R2 (%)	R3 (%)	R4 (%)
1	53	18	7	22
2	71	5	5	19
3	70	4	17	9
4	53	20	15	13
5	60	17	13	10
6	47	8	8	37

Percentages of Use subcategories

Instance of dialogue	U1	U2	U3	U4
1	0	18	71	12
2	0	25	25	50
3	0	33	67	0
4	0	20	80	0

5	0	45	40	15
6	0	47	33	20

Reporting tool

Note: This tool is to be completed for each separate session that was analyzed

1. Metadata

Grade level	
Subject matter	
Specific Topic	
Length of session (in minutes)	
Number of students	
Characteristics of the teacher	
Brief synopsis of what preceded the session and the relevant ideas that had already been addressed in previous sessions	
Learning objectives to be addressed through this session	
Extent of Interrater reliability	

2. Completed ESRU cycles

Frequency of completed ESRU cycles	f =	Frequency of incomplete ESRU cycles (by	f =

		incomplete cycles we intent to refer to the broken cycles that could have been usefully completed – or continued)	

Illustrative excerpt corresponding to a **completed** cycle (along with the coding)	Brief justification of why this is considered a completed ESRU cycle
Illustrative excerpt corresponding to an **incomplete** cycle (along with the coding)	Brief justification of why this is considered an incomplete ESRU cycle

3. Missed opportunities

Each LWG is asked to provide a list of the different types of missed opportunities they have encountered in their data along with a relevant documentation as shown in the table below

Brief description of the type of missed opportunity (label)	Description of the missed opportunity (this should illustrate why it would be appropriate to conceive of this as a missed opportunity)	Evidence of the existence of this type of missed opportunity (this should include the relevant excerpt(s) along with the corresponding coding at the level of the subcategories)	Frequency (frequency at which each type of missed opportunity was encountered in the session under analysis)

Template for Coding Data from Interview with teacher (interactions on the fly)

- Challenges identified by the teacher with respect to the enactment of this assessment method

Challenge	Evidence
Brief description of the challenge (formulated in a generalizable manner)	Evidence (English translation of the interview excerpt) for the teacher reported challenge

- CONSENSUS ON THE INTERPRETATION: Evidence supporting a specific interpretation made by the researcher for instances of incomplete ESRU cycles

Incomplete ESRU cycle	Researcher's interpretation	Confirming evidence from the teacher
Brief description of the incomplete ESRU cycle (English translation of the excerpt from the video transcription accompanied with brief information for the context in which that discussion took place – where does it refer? What preceded this excerpt?)	Brief description offered by the teacher for why the cycle broke (expressed in a generalizable form)	Evidence confirming this interpretation (interview excerpt –English translation)

- DIFFERING INTERPRETATIONS: Alternative interpretations offered by the teacher for specific instances of incomplete ESRU cycles

Incomplete ESRU cycle	Researcher's interpretation	Alternative interpretation offered by the teacher
Brief description of the incomplete ESRU cycle (English translation of the excerpt from the video transcription accompanied with brief information for the context in which that discussion took place – where does it refer? What preceded this excerpt?)	Brief description offered by the teacher for why the cycle broke (expressed in a generalizable form)	Brief Description of alternative interpretation Evidence of this teacher interpretation (interview excerpt –English translation)

- Evidence supporting a specific type of missed opportunity encountered in the data (other than those described as broken ESRU cycles in the previous items)

Missed opportunity	Confirming evidence from the teacher
Brief description of the missed opportunity with brief documentation (this could be copied from the coding tool that researchers will describe and document the missed opportunities they identified)	Evidence confirming this interpretation (interview excerpt –English translation)

- Evidence questioning the interpretation for a specific instance of missed opportunity

Missed opportunity	Evidence from the teacher interview challenging the interpretation that led to the missed opportunity
Brief description of the missed opportunity with brief documentation (this could be copied from the coding tool that researchers will describe and document the missed opportunities they identified)	Evidence challenging this interpretation (interview excerpt –English translation)

Teacher self-reporting tool

This tool is divided in two parts. The first includes a total of 9 items associated with the assessment method "Interactions on the Fly". The items are expressed in a likert scale and are grouped in two categories depending on whether they refer to your perceptions about (a) the assessment method and (b) students' engagement with this assessment method. The second part includes four open-ended questions.

Name: **Date:**

School:

Part A

Use the 4-point scale shown below to indicate the extent to which you agree or disagree with each statement in the table.

1	*2*	*3*	*4*
Not at all			*To a great extent*

1. The assessment method "Interactions on the Fly":				
1.1. was easy to implement	1	2	3	4
1.2. took more time than I expected	1	2	3	4
1.3. was beneficial to the students	1	2	3	4
1.4. was beneficial to myself as the teacher	1	2	3	4
1.5. could produce information that may inform my instructional decisions about your next steps in the unit	1	2	3	4
1.6. is something I use in my daily teaching practice	1	2	3	4

1.7. is consistent with my usual teaching approach in my class	1	2	3	4
1.8. is consistent with the typical teaching practice of teachers in my country	1	2	3	4
2. Students were:				
2.1. used to engage in structured classroom dialogue	1	2	3	4
2.2. keen to contribute to the discussions	1	2	3	4
2.3. in a position to act (in real time) on the feedback they received during the discussion in real time	1	2	3	4

Part B

Please respond to the following four questions.

What are the main strengths and the main weaknesses of the assessment method "Interactions on the Fly"?
What challenges and what opportunities do you identify in enacting this assessment method?
Opportunities: Challenges:
What resources could help teachers overcome these challenges?

What would you have done differently, in terms of the structure and procedures you have employed, if you had the opportunity to repeat the implementation of this assessment method?

Appendix IV: Research designs

Research Design I

Assessment Method: Interactions "on-the-fly"

Focus of the specific research design: In-depth analysis of the enactment of the assessment method "interactions on the fly" in certain situations as well as documentation of relevant challenges.

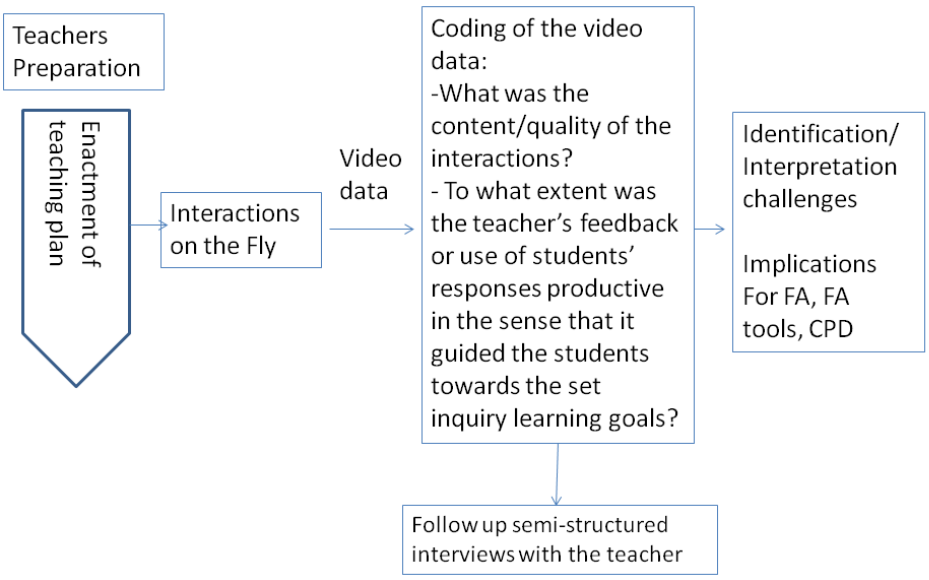
Specific Research Question Associated with the research design

What are the challenges for teachers' use of "interactions on the fly" as a means of formative assessment for promoting a selected inquiry competence?

Corresponding project research goals

What systemic support measures and what tools do teachers need in order to integrate formative assessment of student learning in their classroom practice? **(1.2)**

Illustration of the Research Design



Note: Audio data combined with detailed field notes could be collected instead of classroom video data

Rationale: The activity sequence should be designed so as to incorporate whole-class discussions and/or discussions between smaller groups of students and the teacher. The work and or discussion will be video recorded. The analysis of the data will focus on the content/quality of the interactions in the different situations and the productivity of the teacher's feedback. One aspect of the analysis of the content/quality of the interactions is to identify missed opportunities. (Missed opportunities are defined as instances of the teacher-students conversations in which even though the teacher had a chance to build on students' contributions so as to introduce/elaborate aspects that were among his/her priorities, s/he failed to do so (or refrained from doing so).) Further, teachers are interviewed in a semi-structured fashion on the challenges and opportunities of using this assessment method. The goal of the analysis will be to identify challenges that teachers are confronted with in their attempt to take advantage of "interactions on the fly" as a formative assessment method.

The analysis will focus on the following *analytical questions*:

1. What patterns can we identify in the interactions between the teacher and the students?
2. What factors seem to facilitate or impede teachers' attempt to guide students towards the inquiry learning goals using interactions on the fly?
    a. What are the emergent factors that seem to afford productive teacher feedback?
    b. What are the various types of missed opportunities (as identified in the rationale) encountered in the interactions on the fly; and what are the possible interpretations for why these opportunities were missed by the teacher?
3. What challenges and opportunity do teachers report about their use of "interactions on the fly" as an assessment method?

Scope of the research design/Constraints to be satisfied:

This research design assumes that the teaching to be enacted includes multiple possibilities for teachers to interact with students on the fly defined in the following way (cf. D4.7, p. 18)

- "Interactions on-the-fly [are] informal formative assessment[s] of individual students or small groups of students."
- "On-the-fly assessment cannot be planned beforehand but takes place spontaneously when the teacher recognises good opportunities."

The application of this research design must meet the following *minimal criteria:*

A) The teaching/learning sequence must afford multiple possibilities for teachers to interact spontaneously with students. It must be clear from the implementation scheme where interactions on the fly are expected to occur. For example, will the interactions on the fly occur primarily in whole-class discussions, small group discussions, lab work exercises etc.?
B) The LWG should formulate the explicit learning goals and corresponding levels of attainment. This must not only be used to shape the enacted teaching/learning sequence, but also be used during the video analysis that is performed by two or more researchers from the partner.
C) Partners will be asked to provide an English translation of the specific tool they will be using for collecting and coding data from the teacher for this purpose. In particular, the partner must specify which parts of the teaching/learning sequence will be video-recorded and analyzed.
D) Partners will be asked to describe how the definition of "missed opportunities" applies to their case: how would "missed opportunities" look like in the video data?
E) Teachers will participate in follow-up semi-structured interviews intended to shed light onto aspects of using interactions on the fly as an assessment method.

Note: You will be asked to describe how you will fulfil these criteria in the implementation scheme, and you will be asked to provide a translated version of the assessment tool that was ultimately implemented.

WP5 will provide:

1. A generic coding scheme for the analysis of the video data with respect to
    a. The content of the interactions between the teacher and the student
    b. Reporting and categorizing factors that facilitate or impede teachers' attempt to guide students towards the inquiry learning goals using interactions on the fly
    c. Reporting and categorizing missed opportunities encountered in the interactions on the fly, and the associate interpretations for why these opportunities were missed by the teacher?
2. A generic interview protocol for the semi-structured interviews. *Note:* the individual partner can add to this protocol in order to capture aspects of the teacher's experience that are relevant to the particular context.

WP5 has provided examples meeting the minimal criteria for the competence of investigation.

Additionally this research design assumes that the following criteria are met:

1. The LWG is responsible for ensuring that the teacher's implementation is consistent with the plans of the LWG.
2. The LWG is responsible for supporting the process of collecting the required research data during and after the implementation.
3. The responsibility for the research data collection resides with the researchers of the LWG who are also anticipated to safeguard the inter-rater reliability of the data analysis.

Anticipated output of this research design

At the local level this research design will lead to case studies that will focus on describing specific instances of teachers' attempt to employ "interactions on the fly" as an assessment method. This may include the identification and documentation of challenges associated with the implementation of this assessment method in the context of inquiry-based Teaching/Learning Sequences (TLS) for addressing one of the competences emphasized by the project. This study could bear implications for (a) the practice of formative assessment – e.g., in terms of requirements imposed on students, (b) formative assessment tools (e.g., what tools could support teachers in managing interactions on the fly?) and (c) teachers Continuing Professional Development.

Provided that this research design is implemented by more than one partners it will be possible to also address questions associated with the comparison of the challenges across different situations.

Links to relevant resources:

1. Description of a specific example of how the assessment method "interactions on the fly" can be integrated in an activity sequence about the investigation competence. This is just an example. It could be applied to other teaching/learning sequences either for this or other competence provided that the requirements listed earlier are met.

Data collection:

- Classroom video data or audio data supplemented with detailed field notes
- Teacher interviews following initial coding of the video/audio data

Instruments

Coding tools

Reliability

Research Design II

Assessment Method: Structured Classroom Dialogue

Focus of the specific research design: Exploration of the facility of structured classroom dialogue to serve as an appropriate context for formative assessment.

Specific Research Questions Associated with the research design

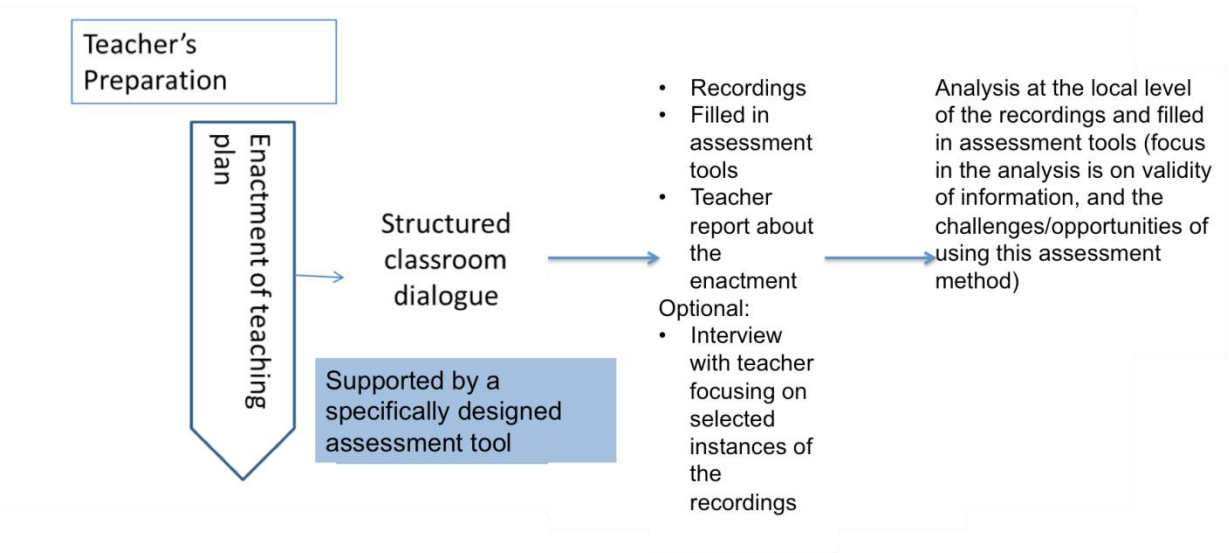
- To what extent can structured classroom dialogue, along with a specially designed assessment tool provide teachers with productive information so as to diagnose students' needs and level of attainment of a selected competence and provide feedback to the students on that basis?

- What are the challenges and opportunities for using structured classroom dialogue, along with a specially designed assessment tool, as a formative assessment method for promoting students' attainment of a selected competence?

Corresponding project research goal

What systemic support measures and what tools do teachers need in order to integrate formative assessment of student learning in their classroom practice? **(1.2)**

Illustration of the Research Design



Rationale: The teaching plan includes a planned classroom event involving structured classroom dialogue. The teacher uses that as a context for collecting information on student attainment of the targeted competence, for interpreting that information and for offering feedback to the students. The structured classroom dialogue will be recorded for future analysis. In addition, a specially designed assessment tool will be used during the structured classroom dialogue. This assessment tool is intended to focus the teacher's attention on the crucial aspects of the dialogue, to capture the core ideas associated with the targeted competence and to facilitate reflective feedback to the students. After the enactment, the teacher writes a report on the enactment along with his/her immediate reflections on the viability of the enactment (a template for these reports will be provided by WP5).

As an *optional* source of data, a partner can interview the teacher by using the stimulated recall interview approach with a focus on exploring selected excerpts of the classroom recording.

WP5 will provide a data-coding tool that will be used at the partner level in order to analyze the collected data.

The analysis will focus on the following *analytical questions*:

4. To what extent do dialogue episodes reveal information on students' level of attainment of the various dimensions of the targeted competence?
    a. To what extent was this information used to provide students with support and formative feedback during and after the structured dialogue?
5. What is the correspondence between the filled-in assessment tools and what actually occurred during the dialogue?
    a. Was there information made available during the dialogue that was not represented in the filled in assessment tool?
    b. What are the differences and similarities between the dynamic feedback during the structured classroom dialogue and the feedback that emerges from the use of the assessment tool and the associate teacher reflection?
6. What challenges and opportunities does the teacher identify for using structured classroom dialogue along with a specially designed assessment tool as a formative assessment method for promoting students' attainment of the learning objective?

Scope of the research design/Constraints to be satisfied:
This research design assumes the implementation of the *structured classroom dialogue* method. This means that the method that is implemented meets two minimal criteria (cf. D4.7, p. 26-7):

1. Before the implementation, some specific rules for the dialogue should be described in detail. These rules formulate how the dialogue process should be structured. Students must be aware of these rules.

2. During the dialogue, students undertake and alternate between different roles – e.g. active presenters, active feedback givers, and active listeners. These roles should be described explicitly.

Note: You will be asked to describe how you intend to reach these criteria in the teaching scheme, and you will be asked to provide a translated version of the rules and the structure of dialogue.

This research design also assumes the existence of a specially designed *assessment tool* that helps the teacher collect information about the students' level of attainment of the given competence in question and also offer feedback. The recording of information could be done during the structured classroom dialogue or immediately afterwards. These assessment tools will have to be developed by the individual LWGs ((where applicable, partners are encouraged to adapt from the examples provided by WP5 – see last section of this document - and collaborate with each other)). There are three minimal criteria for these assessment tools:

1. There needs to be a specific learning objective (and an associated progression of levels of attainment) that clearly corresponds to the competence in question.
2. The assessment tool must allow the teacher to gather information that is demonstrably relevant to the assessment of students' attainment of the learning goal. For example, the tool could be a richly described rubric.
3. The assessment tool must facilitate formative feedback to the students (after reflection) by including a section where the teacher interprets the recorded information and returns written comments to the students (individual or in groups, or for whole class).

Note: You will be asked to describe how you intend to satisfy these criteria in the teaching scheme, and you will be asked to provide a translated version of the assessment tool that was ultimately implemented.

WP5 has provided examples meeting the minimal criteria for the competence of argumentation.

Additionally this research design assumes that the following criteria are met:

1. The LWG is responsible for ensuring that the teacher' implementation is consistent with the plans of the LWG.
2. The LWG is responsible for supporting the process of collecting the required research data during and after the implementation.
3. The responsibility for the research data collection resides with the researchers of the LWG who are also anticipated to safeguard the inter-rater reliability of the data analysis.

Anticipated output of this research design

At the local level this research design will lead to case studies that will identify and document the merits associated with "structured classroom dialogue" as a formative assessment method. It will also serve to document challenges associated with the implementation of this assessment method in inquiry-based Teaching/Learning Sequences for certain competences.

The research output that is produced on the partner-level will be pooled together with the research outputs from other partners using this research design. At the project-level there will be a meta-analysis of this pooled research output.

Given that it is predicted that more than one partner implements this research design, it will be possible in the meta-analysis to address questions associated with the challenges/intricacies associated with this assessment method in different contexts (or with different competences). For example:

- What is the effect of specific types of structures of classroom dialogue on the quality of information about students' attainment of given learning objectives?

Links to relevant resources:

Description of a specific example of how structured classroom dialogue has been integrated in an activity sequence about the argumentation competence situated in a socio-scientific issue. This is just an example. This research design could be applied to other teaching/learning sequences either for this competence or one of the other two competences (investigation or modeling) provided that the requirements listed earlier are met.

Data Collection:

Classroom video data or audio data supplemented with detailed field notes

Instruments

Reliability

Reliability:

Research Design III

Assessment Method: Marking (Grading and Written Comments)

Focus of the specific research design: Investigation of the potential of certain tools to facilitate the effective implementation of the assessment method "Marking (Grading and Written Comments)", by supporting teachers' attempt to interpret students' data, diagnose difficulties/needs and provide them with productive[11] feedback.

Specific Research Questions Associated with the research design

- To what extent can a specially designed assessment tool for *Marking (Grading and Written Comments)* provide teachers with productive information so as to diagnose students' needs and level of attainment of a selected competence and provide feedback to the students on that basis?

- What are the various ways in which students respond to the feedback they receive? Is there a connection between different "responses" and type of feedback?

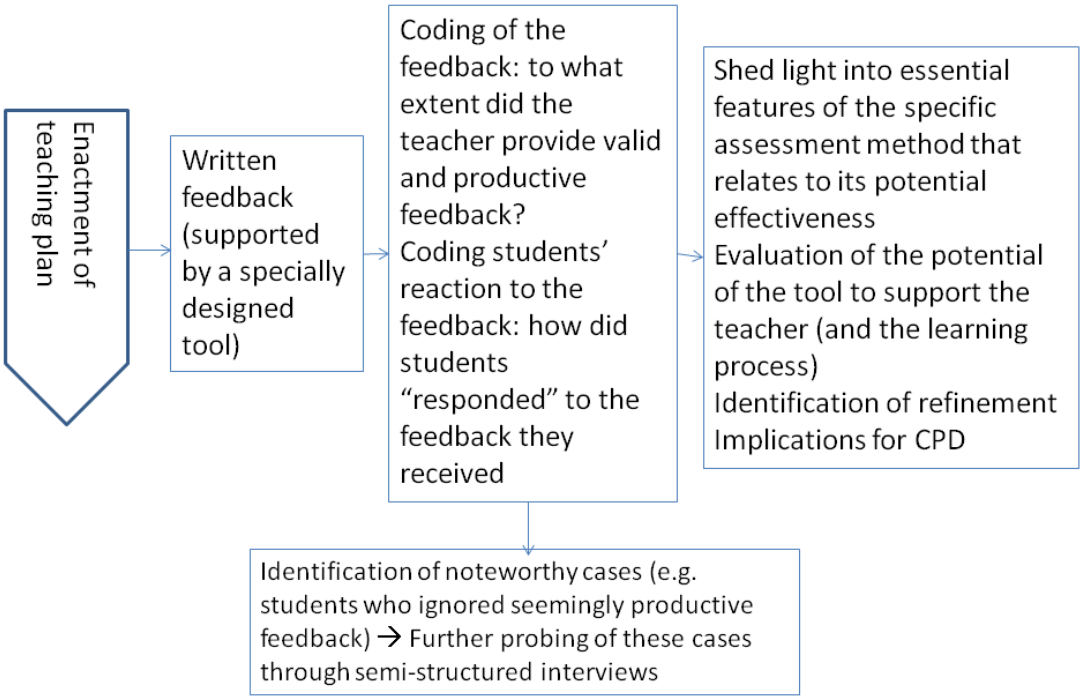
- What are the challenges and opportunities for using *Marking (Grading and Written Comments)*, along with a specially designed assessment tool, as a formative assessment method for promoting students' attainment of a selected competence?

Corresponding project research goals

What systemic support measures and what tools do teachers need in order to integrate formative assessment of student learning in their classroom practice? **(1.2)**

Illustration of the Research Design

[11] Productivity of feedback relates to the extent to which it guides students towards learning goals (it initiates actions that could facilitate learning).



Rationale: At a specific point of the activity sequence, students submit to the teacher certain artefacts they have produced during the enacted teaching, associated with the competence/sub-competence under emphasis (see indicative list of possible types of artefacts at the end of this document). The teacher provides written feedback to each student. For this s/he uses a specially designed tool intended to facilitate his/her attempt to diagnose students' needs or difficulties, with respect to the competence/sub-competence under emphasis, but also their achievements. The feedback provided by the teacher will be coded. Students will then respond in writing to the comments they received and they will also undertake to revise the initial version of their artefact, taking into account the comment they received. The researchers use the data collected (i.e., initial version of student artefact, feedback comments, revised artefact and response to the comments) to evaluate the extent to which this assessment method was implemented in a productive manner and identify and document possible challenges or intricacies.

The analysis will focus on the following *analytical questions*:

7. To what extent did the tool intended to support teacher's attempts to provide students with feedback productively served this purpose?

    a. Did the teacher produce relevant, productive feedback that was well targeted at students' needs? What is the correspondence between the feedback provided by the teacher and the feedback that would be provided by a knowledgeable peer with expertise about the competence/sub-competence under emphasis?

8. How did students respond to the feedback?
    a. What are the various ways in which students responded to the feedback they received?
    b. To what extent did they actually draw on the feedback comments for revising the initial version of their artefacts?
    c. What are the possible interpretations for the instances in which students who were provided with seemingly productive feedback failed to use it (e.g., while revising the initial version of the artefact they had produced)?
    d. What are the possible interpretations for the instances in which students who produced artefacts that addressed the important criteria (connected to the competence at hand) avoided to draw on these criteria (or drew on them in a non-valid manner) in the feedback comments they prepared for their peers?
9. What are the challenges associated with the implementation of the assessment method Marking (Grading and Written Comments)? What obstacles seem to impede its productive enactment and what are possible ways of addressing them?

Scope of the research design/Constraints to be satisfied:

This research design assumes the implementation of the assessment method *Marking (Grading and Written Comments)* (cf. D4.7, p. 21-22). This means that the method that is implemented meets four minimal criteria:

1. As part of the activity sequence students should be producing certain artefacts associated with the competence promoted through the teaching intervention (e.g., an argument in the case of the argumentation competence, a model of a physical phenomenon in the case of the modelling competence or a design for an experimental design – which variable to alter, which variable to keep constant - in the case of the investigation competence). Each partner will be asked to describe the artefact that will be constructed and to demonstrate how that is linked to the relevant competence.
2. Teachers' feedback will be focused on these artefacts. For instance in the case in which the artefact is a student constructed argument, the feedback could be focusing on the extent to which it contains what are considered to be the essential components such as the claim and the data supporting the claim along with a relevant justification. Feedback should be given to individual students.
3. Teacher's attempt to interpret data from students' artefacts and produce feedback will be supported through specially designed templates. These will be developed by the LWGs.
4. Upon receiving comments from the teacher, each student revises the relevant artefact, taking into account the feedback s/he received. In addition to just revising the artefact, students are also explicitly asked to briefly respond, in writing, to the feedback comments they received (e.g., what did the feedback mean to you? What did you learn from the feedback and how did you use it in your work?).

Note: You will be asked to describe how you will fulfil these criteria in the teaching scheme you will use.

This research design also assumes the development and use of certain assessment tools, as follows.

1. The template that will be used by the students to present the initial version of their artefact. This will be what the teachers' will be focusing his/her feedback on.
2. A template that will be used by the students to respond to the comments they received from their teacher.
3. A template that will be used by the students to present/describe the revised version of their artefact, after receiving feedback comments by the teacher.

Notes:

1. These assessment tools will have to be developed by the individual LWGs (where applicable, partners are encouraged to adapt from the examples provided by WP5 – see last section of this document - and collaborate with each other). WP5 has provided examples meeting the minimal criteria for the following competences: investigation, and design in technology.

2. You will be asked to provide a translated version of the assessments tool that you will use.

Additionally this research design assumes that the following criteria are met:

1. The researchers in the LWG will conduct semi-structured follow-up interviews with students who exhibited a noteworthy behaviour (as described in the second analytical question) in terms of acting on the feedback they received.
2. The LWG is responsible to ensure that the teacher's implementation is consistent with the plans of the LWG.
3. The LWG is responsible for supporting the process of collecting the required research data during and after the implementation.
4. The responsibility for the research data collection resides with the researchers of the LWG who are also anticipated to safeguard the inter-rater reliability of the data analysis.

Anticipated output of this research design

At the local level this research design will lead to case studies that will focus on the implementation of Marking (as an assessment method) in a specific situation. This could be focused on documenting intricacies/patterns identified in that situation (e.g., how students responded in the cases in which they were provided with productive feedback? Also, it could focus on possible challenges encountered by the teachers (to what extent did the use of a particular tool supported teacher's attempt to provide productive feedback to the students?)

Provided that this research design will be implemented by more than one partners it will be possible to also address questions associated with the comparison of the challenges across different situations.

Links to relevant resources:

2. Description of a specific example of how Marking (Grading and Written Comments) has been integrated in an activity sequence about the Investigation competence.
3. Description of a specific example of how Marking (Grading and Written Comments) has been integrated in an activity sequence about Design in Technology.

   These two are just examples. They could be replaced by other teaching/learning sequences either for these same competences or for one of the other two competences (argumentation or modelling). However, in each case it is necessary to ensure that the requirements described in this document are satisfied.

Indicative types of artefacts associated with various competences:

- Argumentation: the artefacts could be student constructed arguments and the feedback comments could be referring to the extent to which the arguments contain certain structural elements (e.g. based on Toulmin's model).

- Investigation: the artefacts could be the experimental designs proposed by the students for addressing a specific investigable question (e.g. identification of the variables to change or keep constant). The feedback comments could be focusing on the extent to which the design secures appropriate control of variables and is likely to address the relevant investigable question in a credible manner.

- Modelling: the artefacts could be student-constructed models for specific physical phenomena and the focus of the feedback could be placed on the extent to which these models are characterized by representational, interpretive and predictive capability with respect to the phenomenon of interest.

- Design in Technology: The artefact could be a design product/solution developed through the successful completion of the design process as a response to a specific technological problem. The feedback comments could be focusing on the 'realization stage' of testing the prototype by collecting, analysing, interpreting and representing data.

-Problem solving in Mathematics: the artifact could be a students' solution to a given mathematical problem solving task. The feedback comments in this case could be focusing on the mathematical correctness of the solution given.

Research Design V

Assessment Method: Self & Peer-feedback

Focus of the specific research design: Investigation of the facility of peer feedback to serve as an effective method for formative assessment and exploration of the potential of a specific tool to enhance it.

Specific Research Questions associated with the research design:

- To what extent does the implementation of peer feedback, scaffolded by specific templates,

(a) help students to offer productive feedback on their peers' artifacts[12]?

(b) engage students (as peer-assessees) in the process of using the feedback they received to revise their artifacts?

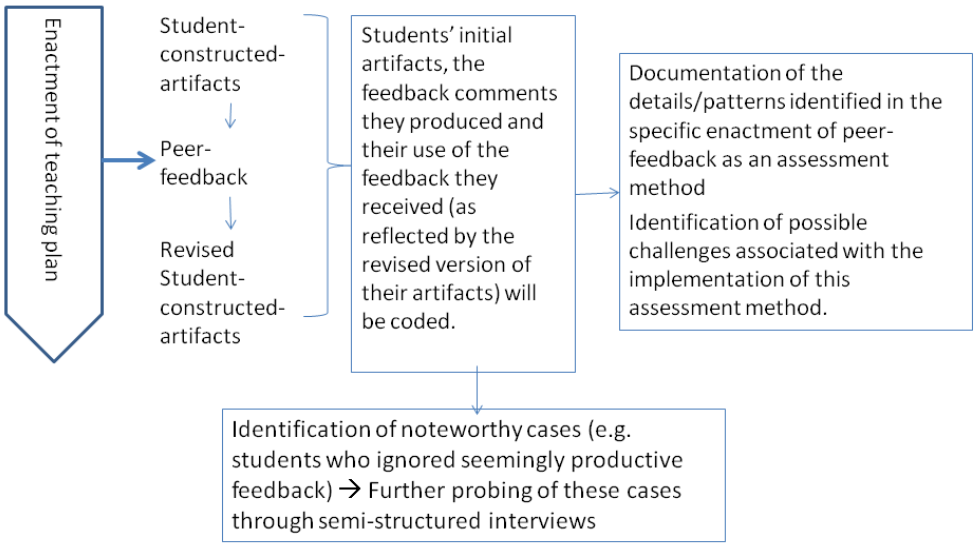
- What challenges become relevant to the implementation of peer feedback as an assessment method?
- How can peer assessment be aligned with the learning goals and be integrated with the teaching sequence

Corresponding project research goals

- What systemic support measures and what tools do teachers need in order to integrate formative assessment of student learning in their classroom practice? (**1.2**)

Illustration of the Research Design

[12] This term refers to student-constructed artifacts that are profoundly connected to the competence of interest. See the list of possible types of artifacts associated with various competences emphasized by AssistMe at the end of this document.



Rationale: At a specific point of the teaching sequence, students engage in the process of peer feedback in a structured manner: each student exchanges an artifact ((see indicative list of possible types of artefacts at the end of this document)). Students will be scaffolded to structure their comments using specified criteria (these will depend on the specific competence/sub-competences under emphasis). Upon receiving comments from a peer, each student undertakes to revise his/her initial argument, accordingly. The researchers use the data from the students' peer feedback (initial artifact, feedback comments and revised artifact) to evaluate the extent to which this assessment was implemented in a productive manner and identify and document possible challenges or intricacies.

The analysis will focus on the following *analytical questions*:

10. To what extent are assessors/assessees engaged productively with the peer-feedback process?

    a. To what extent did students provide relevant, productive feedback to their peers? What is the correspondence between the feedback provided by the students and the feedback that would be provided by a teacher or expert with expertise about the competence/sub-competence under emphasis?

    b. What are the various ways in which students responded to the feedback they received from their peers?

    c. To what extent did they actually draw on the feedback comments for revising the initial version of their artifacts?

    d. What are the possible interpretations for noteworthy behaviors exhibited by students when acting either as peer-assessors or peer-assessees? One instance of such noteworthy behavior refers to the students who were provided with seemingly

productive feedback, though failed to use it while revising the initial version of the artifact they had produced. Another instance, relates to the peer-assessors who were demonstrably in a position to provide relevant, useful feedback though refrained from doing so. A case in point will be the students who produced artifacts that reflected attainment of an aspect of the relevant competence/sub-competence(s) – hence, they were in a position to offer useful feedback, though provided instead very poor or irrelevant feedback to their peers.

11. What challenges emerge in attempts to engage students in the process of peer-feedback? What obstacles seem to impede the productive enactment of the roles of the peer-assessor and peer-assessee and what are possible ways of addressing them?

Scope of the research design/Constraints to be satisfied:

This research design assumes the implementation of the *peer assessment* method (cf. D4.7, p. 23-5). This means that the method that is implemented meets five minimal criteria:

1. There needs to be a specific learning objective (and an associated progression of levels of attainment) that clearly corresponds to the competence in question.
2. As part of the activity sequence students produce certain artifacts associated with the competence promoted through the teaching intervention (see the list of possible types of artefacts associated with various competences emphasized by AssistMe at the end of this document). Each partner will be asked to describe the artifacts that will be constructed and to demonstrate how that is linked to the relevant competence.
3. The implementation of the peer-feedback method will be focused on these artifacts. Students provide written feedback about the relevant artifacts to other students. Partners will need to provide a translated version of the specific procedures that will be followed during the implementation of the peer-feedback.
4. The process of exchanging peer-feedback will be supported through specially designed templates, which will encompass criteria for assessing the specific artifacts. These tools will be developed by the LWGs.
5. Before the implementation of the assessment method the students should be introduced to the roles of the peer-assessor and the peer-assessee. During the implementation students alternate between the two roles (peer-assessor and peer assessee).
6. Upon receiving feedback from his/her peers, each student should be asked to revise the relevant artifact, taking into account the feedback s/he received. In addition to just revising the artifact or to describe how they would revise the artifact or to apply what they have learnt in a new product (e.g. journal notes → feedback → final report), s/he will be also explicitly asked to briefly respond, in writing, to the feedback comments they received.

Notes: You will be asked to describe how you will fulfill these criteria in the teaching scheme. WP5 has provided examples meeting the minimal criteria for the competences of argumentation and modeling .

This research design also assumes the development and use of certain assessment tools, as follows.

1. A template that will be used by the students (peer-assessors) to provide comments to their peers. For instance, this could specify the criteria that should be used by the students for assessing their peers' artifacts.
2. A template that will be used by the students (assessees) to respond to the comments they received from their peers.
3. A template that will be used by the students to present/describe the revised version of their artifact, after receiving feedback comments by their peers.

These assessment tools will have to be developed by the individual LWGs (where applicable, partners are encouraged to adapt from the examples provided by WP5 – see last section of this document - and collaborate with each other)). Partners will be asked to provide a translated version of the assessments tool they will use.

Additionally this research design assumes that the following criteria are met:

1. The researchers in the LWG will conduct semi-structured follow-up interviews with students who exhibited a noteworthy behavior while enacting either the role of the peer-assessor or the peer-assessee.
2. The LWG is responsible for ensuring that the teacher's implementation is consistent with the plans of the LWG.
3. The LWG is responsible for supporting the process of collecting the required research data during and after the implementation.
4. The responsibility for the research data collection resides with the researchers of the LWG who are also anticipated to safeguard the inter-rater reliability of the data analysis.

Anticipated output of this research design

At the local level this research design will lead to case studies that will focus on the implementation of peer feedback (as an assessment method) in a specific situation. This could be focused on documenting intricacies/patterns identified in that situation (e.g., how students responded in the cases in which they were provided with productive feedback?)

Provided that this research design will be implemented by more than one partners it will be possible to also address questions associated with the challenges encountered in different contexts.

Links to relevant resources:

Description of a specific example of how peer-feedback can be integrated in an activity sequence about the argumentation competence. These are just examples. The research design could be applied to other teaching/learning sequences either for this or other competences provided that the constraints associated with this research design are satisfied.

Indicative types of artifact associated with various competences:

- Argumentation: the artifacts could be student constructed arguments and the feedback comments could be referring to the extent to which the arguments contain certain structural elements (e.g. based on Toulmin's model).

- Investigation: the artifacts could be the experimental designs proposed by the students for addressing a specific investigable question (e.g. identification of the variables to change or keep constant). The feedback comments could be focusing on the extent to which the design secures appropriate control of variables and is likely to address the relevant investigable question in a credible manner.

- Modeling: the artifacts could be student-constructed models for specific physical phenomena and the focus of the feedback could be placed on the extent to which these models are characterized by representational, interpretive and predictive capability with respect to the phenomenon of interest.

- Design in Technology: The artifact could be a design product/solution developed through the successful completion of the design process as a response to a specific technological problem. The feedback comments could be focusing on the 'realization stage' of testing the prototype by collecting, analysing, interpreting and representing data.

-Problem solving in Mathematics: the artifact could be a students' solution to a given mathematical problem solving task. The feedback comments in this case could be focusing on the mathematical correctness of the solution given.