

**Towards the implementation of formal formative assessment
in inquiry-based science education in Switzerland**

Inauguraldissertation

zur

Erlangung der Würde eines Doktors der Philosophie
vorgelegt der
Philosophisch-Naturwissenschaftlichen Fakultät
der Universität Basel

von

Regula Grob
aus Wattwil SG

Basel, 2017

Originaldokument gespeichert auf dem Dokumentenserver der Universität Basel
edoc.unibas.ch

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät

auf Antrag von

Prof. Dr. Andreas Wetzel (Fakultätsvertreter)

Prof. Dr. Peter Labudde (Dissertationsleiter)

Prof. Dr. Jens Dolin (Korreferent)

Basel, den 13. 12. 2016

Prof. Dr. Jörg Schibler

List of contents

1	INTRODUCTION.....	1
2	CONTEXT OF THE STUDY	3
3	THEORY.....	4
3.1	THE CONCEPT OF INQUIRY-BASED SCIENCE EDUCATION	4
3.2	THE CONCEPT OF FORMATIVE ASSESSMENT.....	9
3.3	MECHANISMS IN FORMATIVE ASSESSMENT THAT SUPPORT LEARNING	13
3.4	METHODS OF FORMATIVE ASSESSMENT FOR THE CONTEXT OF INQUIRY-BASED SCIENCE EDUCATION	16
3.5	TEACHER CONCEPTS OF AND SELF-EFFICACY IN FORMATIVE ASSESSMENT.....	20
3.6	OBSTACLES TO PUTTING FORMATIVE ASSESSMENT INTO PRACTICE AND MEASURES OF SUPPORT	23
3.7	EFFECTS OF PROGRAMS IMPLEMENTING FORMATIVE ASSESSMENT.....	27
3.8	INQUIRY AND ASSESSMENT IN SWITZERLAND.....	30
4	RESEARCH QUESTIONS.....	38
4.1	A THEORETICAL FRAME FOR INNOVATION IN TEACHING: THE MODEL OF PROFESSIONAL GROWTH	38
4.2	INTRODUCTION OF RESEARCH QUESTIONS.....	40
4.3	USE OF RESULTS FOR GENERATION OF HYPOTHESES	42
5	METHODS	43
5.1	SETTING	43
5.2	PARTICIPANTS.....	45
5.3	DATA COLLECTION	47
5.4	SELECTION OF CASES FOR ANALYSIS	56
5.5	DATA ANALYSIS.....	58
6	ILLUSTRATIVE EXAMPLES OF IMPLEMENTATIONS.....	67
6.1	WRITTEN TEACHER ASSESSMENT AT PRIMARY SCHOOL	67
6.2	PEER-ASSESSMENT AT UPPER SECONDARY SCHOOL	68
6.3	COMBINATION OF PEER-ASSESSMENT AND SELF-ASSESSMENT AT PRIMARY SCHOOL	69
7	RESULTS.....	71
7.1	RESULTS ON RESEARCH QUESTION 1: TEACHERS' UNDERSTANDING OF FORMATIVE ASSESSMENT	73
7.2	RESULTS ON RESEARCH QUESTION 2: DESCRIPTION AND ANALYSIS OF THE TEACHERS' TRIALS	76
7.3	RESULTS ON RESEARCH QUESTION 3: TEACHERS' AND STUDENTS' EVALUATIONS OF THE METHODS TRIALLED	92
7.4	RESULTS ON RESEARCH QUESTION 4: CHANGES IN TEACHERS' UNDERSTANDINGS AND IMPLEMENTATIONS THROUGHOUT THE COLLABORATION IN THE STUDY	112
8	DISCUSSION	128
8.1	DISCUSSION OF RESEARCH QUESTION 1: THE TEACHERS' UNDERSTANDING OF FORMATIVE ASSESSMENT	128
8.2	DISCUSSION OF RESEARCH QUESTION 2: DESCRIPTION AND ANALYSIS OF THE TEACHERS' TRIALS.....	130
8.3	DISCUSSION OF RESEARCH QUESTION 3: TEACHERS' AND STUDENTS' EVALUATIONS OF THE METHODS TRIALLED	137
8.4	DISCUSSION OF RESEARCH QUESTION 4: CHANGES IN TEACHERS' UNDERSTANDINGS AND IMPLEMENTATIONS THROUGHOUT THE COLLABORATION IN THE STUDY	144
8.5	MEASURES OF SUPPORT FOR FORMAL FORMATIVE ASSESSMENT PRACTICES IN SWITZERLAND.....	148
8.6	TEACHERS DEVELOPING THEIR OWN FORMATIVE ASSESSMENT PRACTICES IN THE CONTEXT OF THIS STUDY	154
9	RETROSPECTS AND PROSPECTS.....	156
9.1	AIMS OF THE STUDY	156
9.2	CRITIQUE OF METHODOLOGY	156
9.3	IMPLICATIONS OF THE STUDY	157
9.4	PROSPECTS	157
10	LITERATURE	159

APPENDIX.....	172
A1. TEACHER PROFILE QUESTIONNAIRE.....	172
A2. QUESTION FOR TEACHERS TO DEFINE “FORMATIVE ASSESSMENT”	175
A3. EVALUATION FORM FOR TEACHERS	176
A4. EVALUATION FORM FOR STUDENTS	180
A5. INTERVIEW QUESTIONS FOR TEACHERS.....	182
A6. TOPICS FOR TEACHER GROUP DISCUSSIONS	184
A7. DESCRIPTION OF CASES.....	185
A8. DESCRIPTION OF CATEGORIES FOR RQ 1	192
A9. DESCRIPTION OF CATEGORIES FOR RQ 2	193
A10. DESCRIPTION OF CATEGORIES FOR RQs 3.2 AND 3.5	197
A11. DESCRIPTION OF CATEGORIES FOR RQ 3.3	201
A12. DESCRIPTION OF CATEGORIES FOR RQ 3.6	202
A13. DESCRIPTION OF CATEGORIES FOR RQ 4.5	203

List of tables

TABLE 1: CYCLE LENGTHS FOR FORMATIVE ASSESSMENT (FROM WILIAM, 2010)	12
TABLE 2: ALIGNMENT OF INQUIRY ACTIVITIES AS DEFINED IN BELL ET AL. (2010), THE SCIENCE COMPETENCE MODEL (HARMOs, 2008) AND THE MANNERS OF THINKING, WORKING AND ACTING (D-EDK, 2014) FROM THE CURRICULUM OF THE COMPULSORY SCHOOL LEVELS.	31
TABLE 3: ALIGNMENT OF INQUIRY ACTIVITIES AS DEFINED IN BELL ET AL. (2010) AND THE CURRICULUM FOR THE GYMNASIUM (RLP NACH MAR).	32
TABLE 4: CONNECTIONS BETWEEN RESEARCH QUESTIONS FROM SUB-CHAPTER 4.2 AND THE HYPOTHESES DERIVED FROM THE RESULTS IN SUB-CHAPTERS 8.5 AND 8.6.	42
TABLE 5: PARTICIPANTS OF THE STUDY TEACHING AT PRIMARY SCHOOL (N=9). TEACHERS MARKED WITH AN ASTERISK LEFT THE PROJECT AFTER TWO SEMESTERS.	45
TABLE 6: PARTICIPANTS OF THE STUDY TEACHING AT UPPER SECONDARY SCHOOL (N=11). TEACHERS MARKED WITH TWO ASTERISKS COLLABORATED WITH OLIA TSIVITANIDOU FOR ONE (S8) OR TWO (S9; S10) SEMESTERS. THE TRIALS THAT EMERGED FROM THAT COLLABORATION ARE NOT INCLUDED IN THIS STUDY.	46
TABLE 7: DATA FROM THE TEACHER PROFILE QUESTIONNAIRE	49
TABLE 8: DATA FROM THE WRITTEN DEFINITION TASK.....	50
TABLE 9: DATA FROM THE EVALUATION FORMS FOR TEACHERS.....	51
TABLE 10: TEACHING MATERIALS.	51
TABLE 11: ASSESSED STUDENT ARTEFACTS AND CORRESPONDING FEEDBACK.....	52
TABLE 12: OBSERVATION NOTES FROM LESSONS VISITED.	53
TABLE 13: DATA FROM THE EVALUATION FORMS FOR STUDENTS.	53
TABLE 14: DATA FROM THE INDIVIDUAL INTERVIEWS WITH TEACHERS.....	54
TABLE 15: DATA FROM THE GROUP DISCUSSIONS WITH THE TEACHERS.	55
TABLE 16: OVERVIEW OF CASES.....	57
TABLE 17: CODING FRAME FOR THE FIRST RESEARCH QUESTION.	58
TABLE 18: CODING FRAME FOR RESEARCH QUESTION 2.....	59
TABLE 19: CODING FRAME FOR THE THIRD RESEARCH QUESTION.	61
TABLE 20: CODING FRAME FOR THE FOURTH RESEARCH QUESTION.....	63
TABLE 21: TEACHER PROFILE QUESTIONNAIRES THAT WERE INCLUDED IN THE ANALYSIS.....	63
TABLE 22: OVERVIEW OF SCALES BUILT FROM THE ITEMS OF THE TEACHER PROFILE QUESTIONNAIRE.....	64
TABLE 23: TEACHERS' UNDERSTANDING OF FORMATIVE ASSESSMENT.....	73
TABLE 24: PRIMARY SCHOOL TEACHERS (IN N=18 METHODS TRIALLED IN 14 TRIALS): USABILITY OF METHODS. 4=VERY USEFUL; 1=VERY USELESS.....	92
TABLE 25: UPPER SECONDARY SCHOOL TEACHERS (IN N=22 METHODS TRIALLED IN 20 TRIALS): USABILITY OF METHODS. 4=VERY USEFUL; 1=VERY USELESS.	92
TABLE 26: CODING SYSTEM FOR BENEFITS AND CHALLENGES OF ASSESSMENT METHODS TRIALLED AS PERCEIVED BY THE TEACHERS.	94
TABLE 27: BENEFITS OF WRITTEN TEACHER ASSESSMENT (N=17 TRIALS FROM 12 TEACHERS): CATEGORIES AND SUB-CATEGORIES (LEFT) AS WELL AS PARAPHRASES OF QUOTES (RIGHT).	100
TABLE 28: CHALLENGES OF WRITTEN TEACHER ASSESSMENT (N=17 TRIALS FROM 12 TEACHERS): CATEGORIES AND SUB-CATEGORIES (LEFT) AS WELL AS PARAPHRASES OF QUOTES (RIGHT).	101
TABLE 29: BENEFITS OF PEER-ASSESSMENT AS MENTIONED BY THE TEACHERS (N=17 TRIALS FROM 13 TEACHERS): CATEGORIES AND SUB-CATEGORIES (LEFT) AS WELL AS PARAPHRASES OF QUOTES (RIGHT).	101
TABLE 30: CHALLENGES RELATED TO PEER-ASSESSMENT (N=17 TRIALS FROM 13 TEACHERS): CATEGORIES AND SUB-CATEGORIES (LEFT) AS WELL AS PARAPHRASES OF QUOTES (RIGHT).	103
TABLE 31: BENEFITS OF SELF-ASSESSMENT AS PERCEIVED BY THE TEACHERS (N=6 TRIALS FROM 6 TEACHERS): CATEGORIES AND SUB-CATEGORIES (LEFT) AS WELL AS PARAPHRASES OF QUOTES (RIGHT).....	104
TABLE 32: CHALLENGES OF SELF-ASSESSMENT AS PERCEIVED BY THE TEACHERS (N=6 TRIALS FROM 6 TEACHERS): CATEGORIES AND SUB-CATEGORIES (LEFT) AS WELL AS PARAPHRASES OF QUOTES (RIGHT).	104
TABLE 33: USABILITY OF WRITING COMMENTS IN THE CONTEXT OF PEER-ASSESSMENT AS EVALUATED BY THE STUDENTS. 4=VERY USEFUL; 1=VERY USELESS. AM = ARITHMETIC MEAN; SD = STANDARD DEVIATION. CLASSES 1-3 ARE SUMMARIZED AS ONE SETTING BECAUSE THE RESPECTIVE PEER-ASSESSMENT WAS CONDUCTED BY THE SAME TEACHER IN THE SAME INQUIRY-BASED UNIT WITH THREE DIFFERENT CLASSES.	106
TABLE 34: USABILITY OF COMMENTS FROM PEERS AS EVALUATED BY THE STUDENTS. 4=VERY USEFUL; 1=VERY USELESS. AM = ARITHMETIC MEAN; SD = STANDARD DEVIATION. CLASSES 1-3 ARE SUMMARIZED AS ONE SETTING BECAUSE THE	

RESPECTIVE PEER-ASSESSMENT WAS CONDUCTED BY THE SAME TEACHER IN THE SAME INQUIRY-BASED UNIT WITH THREE DIFFERENT CLASSES.	106
TABLE 35: CODING SYSTEM FOR BENEFITS AND CHALLENGES OF PEER-ASSESSMENT (BOTH ASSESSOR AND ASSESSE ROLES) AS PERCEIVED BY THE STUDENTS.	107
TABLE 36: BENEFITS OF ASSESSING PEERS AS REPORTED BY THE STUDENTS (N=103 STUDENTS FROM 5 CLASSES). CATEGORIES AND SUB-CATEGORIES (LEFT) AS WELL AS PARAPHRASES OF QUOTES (RIGHT).	108
TABLE 37: BENEFITS OF RECEIVING FEEDBACK FROM PEERS AS REPORTED BY THE STUDENTS (N=103 STUDENTS FROM 5 CLASSES). CATEGORIES AND SUB-CATEGORIES (LEFT) AS WELL AS PARAPHRASES OF QUOTES (RIGHT).	108
TABLE 38: CHALLENGES OF ASSESSING PEERS AS REPORTED BY THE STUDENTS (N=103 STUDENTS FROM 5 CLASSES). CATEGORIES AND SUB-CATEGORIES (LEFT) AS WELL AS PARAPHRASES OF QUOTES (RIGHT).	109
TABLE 39: CHALLENGES OF RECEIVING FEEDBACK FROM PEERS AS REPORTED BY THE STUDENTS (N=103 STUDENTS FROM 5 CLASSES). CATEGORIES AND SUB-CATEGORIES (LEFT) AS WELL AS PARAPHRASES OF QUOTES (RIGHT).	110
TABLE 40: ELEMENTS IN THE TEACHERS' WRITTEN DEFINITIONS IN THE THREE ROUNDS OF IMPLEMENTATION. SINCE THE TEACHERS' WERE ASKED ABOUT THEIR UNDERSTANDING OF THE TERM 'FORMATIVE ASSESSMENT' ANONYMOUSLY, THEIR CODES ARE DIFFERENT FROM THE CODES INTRODUCED IN SUB-CHAPTER 5.2.	113
TABLE 41: DESCRIPTIVE STATISTICS OF THE PERSONAL FORMATIVE ASSESSMENT EFFICACY BELIEF SCALE AND ITS ITEMS. MDN=MEDIAN; AM= ARITHMETIC MEAN; SD= STANDARD DEVIATION.	114
TABLE 42: WILCOXON TESTS AND EFFECT SIZES FOR THE PERSONAL FORMATIVE ASSESSMENT EFFICACY BELIEF SCALE AND ITS' ITEMS. SIGNIFICANCE: *P<0.05; **P<0.01; COHEN'S D: 0.2 <D≤ 0.5 REPRESENTS SMALL EFFECT SIZE; 0.5 <D≤ 0.8 REPRESENTS MEDIUM EFFECT SIZE; 0.8 <D REPRESENTS LARGE EFFECT SIZE (COHEN, 1988).	115
TABLE 43: THE TEACHERS' IMPLEMENTATIONS DISPLAYED OVER TIME. CRITERIA 1)– 4) SPELLED OUT ABOVE.	116
TABLE 44: DESCRIPTIVE STATISTICS OF THE FREQUENCY SCALES. MDN=MEDIAN.	117
TABLE 45: WILCOXON TESTS AND EFFECT SIZES FOR THE FREQUENCY SCALES. SIGNIFICANCE: *P<0.05; **P<0.01.	117
TABLE 46: DESCRIPTIVE STATISTICS OF THE IMPORTANCE SCALES. MDN=MEDIAN; AM= ARITHMETIC MEAN; SD= STANDARD DEVIATION.	118
TABLE 47: WILCOXON TESTS AND EFFECT SIZES FOR THE IMPORTANCE SCALES. SIGNIFICANCE: *P<0.05; **P<0.01; COHEN'S D: 0.2 <D≤ 0.5 REPRESENTS SMALL EFFECT SIZE; 0.5 <D≤ 0.8 REPRESENTS MEDIUM EFFECT SIZE; 0.8 <D REPRESENTS LARGE EFFECT SIZE (COHEN, 1988).	118
TABLE 48: BENEFITS AND CHALLENGES PERCEIVED BY THE TEACHERS IN THE STUDY THROUGHOUT THE THREE ROUNDS OF IMPLEMENTATION.	119
TABLE 49: VARIABILITY OF IMPLEMENTATIONS WITHIN TEACHERS.	122
TABLE 50: OVERLAPS.	122
TABLE 51: OVERLAP OF FIRST TWO SUBSEQUENT TRIALS PER TEACHER GROUPED BY DIFFERENT VARIABLES.	123
TABLE 52: DESCRIPTIVE STATISTICS AND SIGNIFICANCE TESTS OF THE OVERLAPS, GROUPED BY SCHOOL LEVEL. MDN=MEDIAN; AM= ARITHMETIC MEAN; SD= STANDARD DEVIATION. DIFFERENCES BETWEEN THE TRIALS FROM THE TWO SCHOOL LEVELS TESTED BY MANN U WHITNEY (*P<0.05; **P<0.01) AND EFFECT SIZES WITH COHEN'S D; 0.2 <D≤ 0.5 REPRESENTS SMALL EFFECT SIZE; 0.5 <D≤ 0.8 REPRESENTS MEDIUM EFFECT SIZE; 0.8 <D REPRESENTS LARGE EFFECT SIZE (COHEN, 1988).	124
TABLE 53: DESCRIPTIVE STATISTICS AND SIGNIFICANCE TESTS OF THE OVERLAPS, GROUPED BY TEACHERS' GENDER. MDN=MEDIAN; AM= ARITHMETIC MEAN; SD= STANDARD DEVIATION. DIFFERENCES BETWEEN THE TRIALS FROM THE TWO SCHOOL LEVELS TESTED BY MANN U WHITNEY (*P<0.05; **P<0.01) AND EFFECT SIZES WITH COHEN'S D; 0.2 <D≤ 0.5 REPRESENTS SMALL EFFECT SIZE; 0.5 <D≤ 0.8 REPRESENTS MEDIUM EFFECT SIZE; 0.8 <D REPRESENTS LARGE EFFECT SIZE (COHEN, 1988).	125
TABLE 54: DESCRIPTIVE STATISTICS AND SIGNIFICANCE TESTS OF THE OVERLAPS, GROUPED BY TEACHING EXPERIENCE. MDN=MEDIAN; AM= ARITHMETIC MEAN; SD= STANDARD DEVIATION. DIFFERENCES BETWEEN THE OVERLAPS OF THE TWO TEACHER GROUPS TESTED BY MANN U WHITNEY (*P<0.05; **P<0.01) AND EFFECT SIZES WITH COHEN'S D; 0.2 <D≤ 0.5 REPRESENTS SMALL EFFECT SIZE; 0.5 <D≤ 0.8 REPRESENTS MEDIUM EFFECT SIZE; 0.8 <D REPRESENTS LARGE EFFECT SIZE (COHEN, 1988).	126
TABLE 55: APPROACHES TO FORMAL FORMATIVE ASSESSMENT AT THE TWO SCHOOL LEVELS INVESTIGATED.	149
TABLE 56: CHARACTERISTIC FEATURES OF THE TRIALS IN THE STUDY.	151

List of figures

FIGURE 1: WORK PACKAGES OF THE ASSIST-ME PROJECT. FROM DOLIN (2012).	3
FIGURE 2: EXAMPLE OF A CYCLIC REPRESENTATION OF SCIENTIFIC INQUIRY (BYBEE, 1997).	5
FIGURE 3: FORMATIVE AND SUMMATIVE ASSESSMENT (BASED ON HARLEN, 2013).	9
FIGURE 4: MODEL OF PROFESSIONAL GROWTH. FROM CLARKE & HOLLINGSWORTH, 2002, P. 951.....	38
FIGURE 5: OVERVIEW OF DATA COLLECTION.	47
FIGURE 6: DATA COLLECTED IN EVERY ROUND OF IMPLEMENTATION.....	48
FIGURE 7: EXAMPLE OF A STUDENT'S CHICK JOURNAL WITH THE TEACHER'S COMMENTS (TEACHER'S COMMENTS CIRCLED IN BLACK).	67
FIGURE 8: STRUCTURING AID FOR PEER-ASSESSMENT AS DEVELOPED BY A TEACHER.	68
FIGURE 9: FORM FOR SELF-ASSESSMENT (UPPER PART) AND PEER-ASSESSMENT (LOWER PART) ON THE EXPLORATION OF A SOIL PROFILE AND THE WORK IN THE STUDENT GROUP.	69
FIGURE 10: RELATIONS BETWEEN THE MODEL OF PROFESSIONAL GROWTH (CLARKE & HOLLINGSWORTH, 2002), AND THE RESULTS ON THE FOUR RESEARCH QUESTIONS IN THIS STUDY.	71
FIGURE 11: DIMENSIONS OF OPENNESS (PRIEMER, 2011) IN THE UNITS OF THE PRIMARY AND THE UPPER SECONDARY SCHOOL TEACHERS.	76
FIGURE 12: NUMBER OF OPEN DIMENSIONS (PRIEMER, 2011) PER TRIAL IN THE UNITS OF THE PRIMARY AND THE UPPER SECONDARY SCHOOL TEACHERS.	77
FIGURE 13: INQUIRY ACTIVITIES (BELL ET AL., 2010) ENACTED IN THE UNITS AT PRIMARY AND AT UPPER SECONDARY SCHOOL.	77
FIGURE 14: NUMBER OF INQUIRY ACTIVITIES PER TRIAL AT PRIMARY AND AT UPPER SECONDARY SCHOOL.....	78
FIGURE 15: DOMAIN-SPECIFIC COMPETENCES (SEE SECTION 3.1.2) ASSESSED.....	78
FIGURE 16: NUMBER OF DOMAIN-SPECIFIC COMPETENCES ASSESSED PER TRIAL AT PRIMARY AND AT UPPER SECONDARY SCHOOL LEVEL.....	79
FIGURE 17: TRANSVERSAL COMPETENCES (SEE SECTION 3.1.2) ASSESSED IN THE STUDY.....	80
FIGURE 18: NUMBER OF TRANSVERSAL COMPETENCES ASSESSED PER TRIAL AT PRIMARY AND AT UPPER SECONDARY SCHOOL LEVEL.....	80
FIGURE 19: INTRODUCTION OF ASSESSMENT CRITERIA.	82
FIGURE 20: DATA SOURCES FOR DIAGNOSIS.	82
FIGURE 21: NUMBER OF DATA SOURCES PER TRIAL AT PRIMARY AND AT UPPER SECONDARY SCHOOL LEVEL.	83
FIGURE 22: ASSESSMENT METHODS.	84
FIGURE 23: DOMAIN-SPECIFIC AND TRANSVERSAL COMPETENCES ASSESSED BY THE DIFFERENT ASSESSMENT METHODS.	85
FIGURE 24: MEANS OF ENGAGING WITH THE FEEDBACK IN THE PRIMARY- AND IN THE UPPER SECONDARY SCHOOL TRIALS....	86
FIGURE 25: MEANS OF ENGAGING WITH THE FEEDBACK DIFFERENTIATED BY COMPETENCE.	86
FIGURE 26: CYCLE LENGTH OF THE TRIALS.	87
FIGURE 27: CYCLE LENGTH OF THE TRIALS DIFFERENTIATED BY COMPETENCES.	88
FIGURE 28: PROBLEMS IN THE TRIALS.....	90
FIGURE 29: BENEFIT SUB-CATEGORIES AS MENTIONED BY THE TEACHERS IN THE STUDY.	98
FIGURE 30: CHALLENGE SUB-CATEGORIES MENTIONED BY THE TEACHERS IN THE STUDY.	99

1 Introduction

Inquiry has been an important part of science educational theory and practice for the last decades. It is usually defined as a set of activities that involves raising questions, planning an experiment or an investigation to answer the questions, conducting the respective actions and collecting data, analysing and interpreting these data. As in other competence-oriented approaches to teaching and learning, the appropriate support and assessment of the students' competences has been much debated in the context of inquiry-based science education.

One way to support and assess students in their learning is formative assessment. The concept is also known as 'assessment for learning' which means that the information on the students' levels of achievement is not used for grading but for planning the next steps in teaching and learning. In that sense, formative assessment is prospectively accompanying learning rather than retrospectively taking stock of the learning success as is summative assessment.

The use of formative assessment methods as a means of support for students' learning is promoted in national (e.g. *Lehrplan 21*, D-EDK, 2014) and international position papers and reports (e.g. OECD 2005; 2013) because of the large positive effect of formative assessment on student learning (e.g. Black & Wiliam, 1998; Hattie, 2009) on the one hand and as a countermovement to large-scale summative assessment (e.g. Harlen, 2007; 2013) on the other hand. However, in the Swiss teaching practice, formative assessment, particularly formal formative assessment methods which involve a certain degree of pre-definition, planning, and formality, are not widely used nor researched.

The implementation of a relatively uncommonly used approach (such as formal formative assessment) into regular teaching practice is not an easy endeavour (e.g. Black & Atkin, 1996; Furtak et al., 2008; Smith & Gorard, 2005; Tierney, 2006). Therefore, the focus of this study is on exploring possibilities and challenges in the implementation of formal formative assessment methods in the context of inquiry-based science education in Switzerland. Since the quality of formative assessment rests to a high degree on the strategies teachers use to gain evidence of student learning and on the use of this evidence to shape subsequent instruction and learning (Bell & Cowie, 2001; Heritage, 2010; Ruiz-Primo, Furtak, Ayala, Yin, & Shavelson, 2010), the emphasis is on the teacher perspective.

In the study, twenty teachers explored and trialled formative assessment methods in their inquiry-based science teaching at primary and at upper secondary school level in Switzerland over the course of three semesters. The formative assessment methods were written teacher assessment, peer-assessment, and self-assessment – so the wide spectrum of methods was reduced to three formal approaches.

The questions explored are:

- 1) What different understandings of the term 'formative assessment' do the teachers have?
In this question, the different views of what 'formative assessment' means according to the teachers collaborating in the study are explored.
- 2) How do the teachers use the methods in their inquiry teaching?
In this question, it is (2.1) explored what inquiry context the teachers use to trial the formative assessment methods, (2.2) how the teachers put the formative assessment methods into practice, and (2.3) what aspects can go wrong in the trials.
- 3) What benefits and challenges do the teachers and the students perceive regarding the formative assessment methods trialled?
The first part of this question focusses on the teachers collaborating in the study: It is (3.1) explored how useful the teachers perceive the different assessment methods for their school levels, (3.2) what benefits and challenges the teachers mention regarding the different assessment methods, (3.3) what means of support the teachers would wish for to enhance their formative assessment practices.
In the second part of the question, the focus is on upper secondary school students who trialled peer-assessment: It is (3.4) investigated how useful the students perceive peer-assessment at their school

level, (3.5) what benefits and challenges the students mention regarding peer-assessment, (3.6) what means of support the students would wish for when assessing their peers.

- 4) How do the teachers' understanding of formative assessment and their trials of formative assessment methods change throughout the three semesters of collaboration?

In this question, different types of changes throughout the study are focussed on: It is (4.1) explored how the teachers' understandings of the term 'formative assessment' change throughout the study; (4.2) how the teachers' formative assessment self-efficacy changes throughout the three semesters of collaboration; (4.3) how the teachers' trials change throughout the three semesters of collaboration; (4.4) what changes can be recognized in the importance, benefits and challenges perceived throughout the study; (4.5) what support mechanisms in the study the teachers perceive; and (4.6) what implementation behaviours the different teachers in the study show.

This is an explorative study researching the conditions for an implementation of formal formative assessment methods to enhance the students' inquiry learning in the educational context of Switzerland. The results provide the grounds for two sets of hypotheses: The first set will concentrate on the opportunities and challenges of such an implementation as well as on potential measures of support. For the second set of hypotheses, the teacher collaboration in the study will be interpreted as a small-scale implementation of formative assessment. Hypotheses on the mechanisms and the outcomes of this collaboration will be deduced.

The study is integrated in the larger ASSIST-ME project. Details on that international project and its relation with the study in Switzerland can be found in chapter 2. In chapter 3, the theoretical background on inquiry-based science education, on formative assessment and on formal formative assessment methods as well as on the implementation of such methods will be summarized. An introduction on the educational context of Switzerland with emphasis on the status of inquiry in science education and on the assessment situation will also be provided. Chapter 4 will introduce the research questions. In chapter 5, the details on the design of the study can be found: A description of the setting, the participants, the data collection and -analysis. Chapter 6 will outline three illustrative examples of trials along with statements from the teachers on these in order to provide an impression of what the trials looked like. The results on the four research questions will be presented in chapter 7, the interpretation and discussion of these results along with the hypotheses mentioned in the above paragraph in chapter 8. The prospects will be summarized in chapter 9.

- The grey dotted line to the left of the text signalizes an introduction to the subsequent sub-chapter and its relevance for the study. In chapter 3, the theory part, in the beginning of every sub-chapter (second-level headings; e.g. 3.1), this meta-text will introduce the purpose of the subsequent sub-chapter with respect to the study.

The grey bar to the left of the text signalizes a concluding summary: In chapter 3, the theory part, at the end of each section (third level headings; e.g. 3.1.1), a summary will be provided and the implications for this study will be indicated. The sections of chapter 7 in the results part (third level headings; e.g. 7.1.1) will also end with summaries. In the discussion part, the sections of sub-chapters 8.5 and 8.6 will end with hypotheses derived from the exploratory results. These will be marked with a grey bar, too.

2 Context of the study

This study was embedded in an international project focussing on assessment in inquiry: ASSIST-ME (Assess Inquiry in Science, Technology and Mathematics Education). It was a collaborative project with ten partner institutes, led by University of Copenhagen representatives. The project's duration was January 2013 to December 2016. The ASSIST-ME project as a whole covered a wide range of school levels and subjects. In Switzerland, the focus was on science education at primary school and on biology, chemistry and physics education at upper secondary school level.

According to its proposal, ASSIST-ME had two aims (Dolin, 2012):

- The development and implementation of formative and summative assessment methods which are usable in inquiry-based education in science, technology and mathematics.
- The elaboration of guidelines for policy makers and other stakeholders for ensuring that assessment enhances learning in science, technology and mathematics education.

In order to reach these aims, sets of assessment methods and competences for assessment were selected at project level. From these sets, every country chose certain subsets of assessment methods and competences which were trialled at certain school levels and subjects. In Switzerland, the focus was on the following assessment methods: Self- and peer-assessment as well as written teacher assessment. In terms of competences, the focus in Switzerland was on the 'investigations in science' competence and its sub-competences. However, in this study, the conceptualisation of the investigation competence follows literature that relates more closely to the understanding of inquiry-based education in Switzerland as reflected in the 'basic competences for science education' (*Grundkompetenzen für die Naturwissenschaften*; EDK, 2011). More details can be found in sub-chapter 3.1 (Understanding of inquiry-based science education and competences ascribed to it) and sub-chapter 3.8 (Situation in Switzerland).

The project was divided into three phases which are displayed in Figure 1: In the first phase, the existing literature on formative assessment was synthesized. Concurrently, the educational systems of the participating countries were characterized. Parts of this work can be found in sub-chapters 3.2 – 3.4 (Formative assessment) and 3.8 (Situation in Switzerland). In the second phase of the ASSIST-ME project, assessment methods and competences were selected and trialled in the different countries. Results from the trials in Switzerland can be found in chapter 7. Phase 3 involved the transformation of the results into national contexts and a number of dissemination activities. These are not covered in this study.

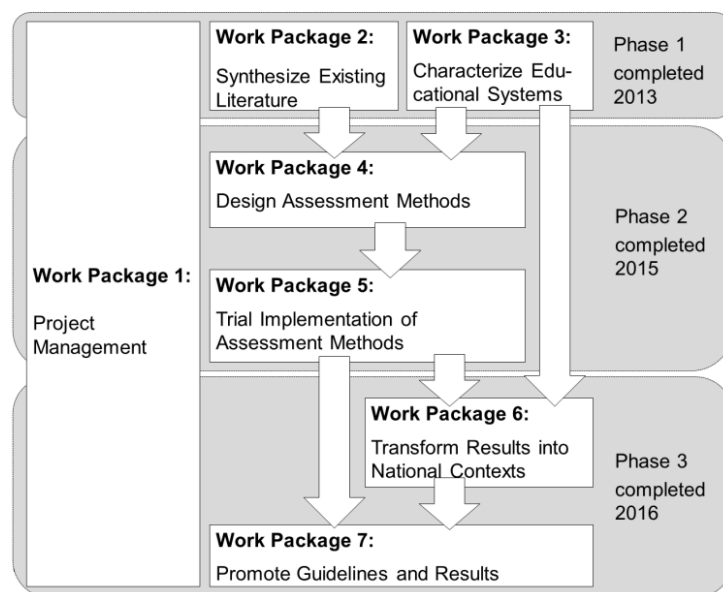


Figure 1: Work packages of the ASSIST-ME project. From Dolin (2012).

3 Theory

This chapter will firstly introduce the concept of inquiry-based science education (sub-chapter 3.1). Afterwards, the concept of formative assessment (sub-chapter 3.2), its mechanisms that support student learning (sub-chapter 3.3) and concrete methods for the context of inquiry (sub-chapter 3.4) will be summarized. The next sub-chapters will focus on the implementation of formative assessment in teaching practice: Aspects of teacher motivation and self-efficacy (sub-chapter 3.5); barriers and challenges reported in the international literature (sub-chapter 3.6); and selected initiatives for implementation (sub-chapter 3.7). The last part of the theory will lay out the educational context of Switzerland with a focus on inquiry-based education and assessment (sub-chapter 3.8).

3.1 The concept of inquiry-based science education

Inquiry has been “a distinguishing feature of innovative science education programs since the 1960s science curriculum reform movement” (Duschl, 2003, p. 41). The US-National Science Education Standards, published in 1996 by the National Research Council, can be seen as a milestone the implementation of inquiry-based science education in theory and practice (Furtak et al., 2012). In these standards, inquiry was described in view of two perspectives (Bybee, 2000; Hofstein, Navon, Kipnis & Mamlok-Naaman 2005; Lunetta, 1998) which still exist in the Next Generation Science Standards (NGSS Lead States, 2013; NRC, 2011): Firstly, the abilities necessary to perform scientific inquiry (such as making observations, posing questions, using tools to gather, analyse and interpret data and communicating the results), which is described as “inquiry as means [...] and refers to inquiry as an instructional approach intended to help students develop understandings of science content” (Abd El Khalick et al., 2004; p. 398). The second perspective is the understanding of scientific inquiry as skills used by scientists which is called “inquiry as ends [...] and refers to inquiry as an instructional outcome: Students learn to do inquiry in the context of science content and develop epistemological understanding about NoS <Nature of Science> and the development of scientific knowledge, as well as relevant inquiry skills” (Abd El Khalick et al., 2004, p. 398).

Inquiry-based science education is an umbrella term subsuming a wide range of approaches to teaching and learning, such as inquiry-based teaching and learning, authentic inquiry, model-based inquiry, modelling and argumentation, project-based science, hands-on science, and constructivist science (Furtak, Seidel, Iverson & Briggs, 2012). Consequently, the characteristics of inquiry-based education vary between different authors (Bell, Urhahn, Schanze & Ploetzner, 2010; Bybee, 2000; Furtak et al., 2012). However, from the many definitions of inquiry-based science education, a number of features to operationalize the term can be deduced: Firstly, inquiry is often described as a set of research-type activities such as investigating phenomena, collecting and interpreting data, communicating and reasoning (e.g. Bybee, 1997). Secondly, inquiry-based science education is typically associated with competence orientation (e.g. Abd El Khalick et al., 2004). Thirdly, inquiry activities contain a certain degree of freedom. That means that not all aspects are pre-defined but that some decisions are left for the students to take (e.g. Priemer, 2011). Finally, inquiry-based science education is rooted in constructivism: the students takes the active parts whereas the teacher acts as a coach to support the students (e.g. Hinrichsen & Jarrett, 1999).

The subsequent sections will focus on how inquiry-based science education can be characterized on a practical classroom level: The key features of inquiry will be introduced. These features are inquiry activities and procedural character of inquiries in science education (section 3.1.1); domain-specific and transversal competences ascribed to scientific inquiry (section 3.1.2); and the openness in the context of scientific inquiry (section 3.1.3). As a transition to the subsequent sub-chapters on formative assessment, the last section will focus on the student-oriented nature of inquiry-based science education (section 3.1.4).

The aim of this sub-chapter is to introduce models and definitions which are tangible enough to deduce categories for the empirical part of this study. These categories will be used to describe the inquiry aspects of the teachers’ trials which are investigated as part of the research questions.

3.1.1 Inquiry activities and procedural character of inquiries

As laid out above, scientific inquiry has often been conceptualized as a set of activities. These activities do not need to follow each other one by one. Instead, it is typical that an inquiry process includes revisions of certain steps and loops back to earlier steps (Artigue & Baptist, 2012). Furthermore, an inquiry does not really end after the last step (typically the conclusions): Instead, the conclusions lead to more questions and hypotheses (e.g. White & Frederiksen, 1998). Some authors therefore represent scientific inquiry in the form of a cycle: Bybee (1997), for example, conceptualizes scientific inquiry as a cycle of five activities. It consists of the 5E that represent engage, explore, explain, elaborate, and evaluate as displayed in Figure 2.

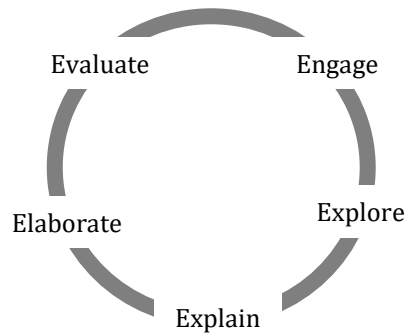


Figure 2: Example of a cyclic representation of scientific inquiry (Bybee, 1997).

In the above conceptualisation by Bybee (1997), the activities that define an inquiry are formulated in a rather abstract way. A more tangible description of activities is suggested by Bell et al. (2010) in their ‘nine main processes of inquiry learning’. The description is based on the analysis of various papers on inquiry processes and terminology (Cuevas, Lee, Hart & Deaktor, 2005; Friedler, Nachmias & Linn, 1990; Gijlers & de Jong, 2005; Harms, Mayer, Hammann, Bayrhuber & Kattmann, 2004; Löhner, van Joolingen, Savelsbergh & van Hout-Wolters, 2005; National Research Council, 1996; Schecker, Fischer & Wiesner, 2004; Schwarz & White, 2005; Singer, Marx, Krajcik & Chambers, 2000; Windschitl, 2004). The nine processes from Bell et al. (2010) will be described below:

- *Orienting and asking questions* are usually at the beginning of an inquiry. Students observe scientific phenomena which catch their curiosity. The authors of the paper (Bell et al., 2010) mention that the process of developing a relevant, investigable question is particularly difficult and may take more than one attempt. This illustrates the cyclic nature of an inquiry which was mentioned above.
- *Hypothesis generation* describes the formulation of relations between different variables (de Jong & Njoo, 1992).
- *Planning* involves, firstly, designing an experiment in order to test the hypothesis, and secondly, to select suitable instruments (Harms et al., 2004).
- *Investigation* represents the empirical part of an inquiry. It includes using instruments to collect data, conducting experiments, and structuring the data pool (Harms et al., 2004).
- *Analysis and interpretation* of the data collected are the basis for empirical claims and arguments. They also serve as the starting point for the development of models (Windschitl, 2004).
- *Model* can be described as “building a cohering whole of objects and relations in order to represent a target area of reality, to reproduce observations from this area, to predict developments, or even to affect developments in this area” (Bell et al., 2010, p. 356). This definition implicitly includes a variety of formats of models such as crafted objects, sketches, mathematical models, or software models.
- *Conclusion and evaluation* describe the extraction of results from an inquiry. Conclusions can be drawn from data and they may include the comparison with theories or other experiments (Harms et al., 2004). Evaluation is a more reflective process that helps students to judge their own research as well as to understand the nature of inquiry (White & Frederiksen, 1998).
- *Communication* describes the collaborative element of an inquiry. It encompasses all other processes, beginning with the development of the research question and ending with the presentation of the results.

- *Prediction* “is a statement about the value(s) of one or more dependent variables under the influence of one or more independent variables” (de Jong & van Joolingen, 1998, p. 189). This last category illustrates a characteristic of inquiry: New questions and hypotheses arise from the conclusions.

In accordance with Bybee (1997) as introduced above, Bell et al. (2010) state that the order of these nine main processes of inquiry learning is not fixed but that “students may go through the processes in the order needed and return to them if necessary” (Bell et al., 2010, p.353).

A distinct feature of inquiry in science education is that inquiry is considered a procedure consisting of several activities. The formulation of these activities varies between authors from more concrete activities (as in Bell et al., 2010) to more abstract formulations (as in Bybee, 1997). The order of these activities is not linear but may contain loops back to a previous activity and is considered cyclic by many authors.

For the purpose of this study which will investigate inquiry units from Swiss classrooms, the conceptualisation of inquiry from Bell et al. (2010) was chosen: Firstly, the activities in this conceptualisation are formulated in a concrete way which is easily applicable to classroom situations. Secondly, the authors have a German background, and this cultural influence results in a focus on the investigative, hands-on parts of an inquiry. This is in close alignment with the frame provided by Swiss curricula such as *Lehrplan 21* and with Swiss teaching practice (more details in sub-chapter 3.8).

3.1.2 Domain-specific and transversal competences ascribed to inquiry

There is no uniform understanding of what the term ‘competence’ means. Hartig, Klieme & Leutner (2008) start from the baseline that “competences can be conceptualized as complex ability constructs that are closely related to performance in real-life situations” (Hartig et al., 2008, Preface). The authors develop a working definition of the term ‘competence’ for use in the context of educational assessment which is similar to the PISA framework. They define competences as “context-specific cognitive dispositions that are acquired by learning and needed for successfully cope with certain situations or tasks in specific domains” (Hartig et al., 2008, p. 9).

Based on this fundamental understanding, most completed and ongoing EU-projects that focus on inquiry (e.g. Mind the Gap, S-TEAM, ESTABLISH and Fibonacci) conceptualize inquiry-based science education as a distinct set of activities and at the same time also consider these activities inquiry competences. For the purpose of this study, the conceptualisation from Bell et al. (2010), introduced in section 3.1.1, will be taken as the basic set of activities that are considered inquiry competences. Therefore, the domain-specific competences ascribed to inquiry-based science education in this study are: Orienting and asking questions, hypothesis generation, planning, investigation, analysis and interpretation, model, conclusion and evaluation, communication, and prediction.

In addition to the above-mentioned domain-specific competences, transversal competences are ascribed to inquiry, too (e.g. Welch, Klopfer, Aikenhead & Robinson, 1981; Zachos, Hick, Doane & Sargent, 2000). Transversal competences cannot be assigned to a particular discipline and are therefore transversal to the traditional structure of scholar disciplines (Grob & Maag Merki, 2001). Terms which, to some extent, overlap with transversal competences are cross curricular competences, key competences (*in German Schlüsselkompetenzen*; Grob & Maag Merki, 2001) as well as 21st century skills and life skills.

The DeSeCo project defined a framework that should guide assessment beyond domain-specific knowledge and skills. The transversal competences defined in this framework are expected to be relevant “for a successful life and a well-functioning society” (OECD, 2005b, p. 4). There is no direct statement that the competences from this framework can be fostered by inquiry. Nevertheless, since the document is recent and the framework is broad enough to capture different understandings of transversal competences as employed in Switzerland, the report from the OECD will form the basis for the conceptualisation of transversal

competences in this study. The competences are classified in three groups which will be introduced below (OECD, 2005b):

- “Use tools interactively including the abilities to use language, symbols and text interactively, to use knowledge and information interactively, and to use technology interactively” (OECD, 2005b, p.10).
- “Interact in heterogeneous groups including the abilities to relate well to others, to co-operate, work in teams, and to manage and resolve conflicts” (OECD, 2005b, p.12).
- “Act autonomously including the abilities to act within the big picture, to form and conduct life plans and personal projects, to defend and assert rights, interests, limits and needs” (OECD, 2005b, p.14).

A distinct feature of inquiry in science education is that both domain-specific and transversal competences are ascribed to inquiry. In a number of projects, the domain-specific competences correspond to the activities that are used to define inquiry (see chapter 3.1.1). The respective transversal competences are more difficult to decide upon. A frequently used framework of transversal competences was defined in the DeSeCo project (OECD, 2005b) and contains three categories. This framework is, however, not directly linked to inquiry.

For the purpose of this study, the domain-specific competences ascribed to inquiry-based science education are correspondent to the activities from Bell et al. (2010) which were used to conceptualise inquiry in chapter 3.1.1. This follows the procedure of many preceding projects. For the transversal competences, the general model published by the OECD (2005b) is taken because it is relatively recent and broad enough to enclose the many and various transversal learning goals focussed on in Swiss classrooms.

3.1.3 Openness of student activities in the context of scientific inquiry

A key feature of scientific inquiry is the openness of the student activities. Openness, in this context, means that inquiry-based education is not entirely pre-defined but involves the freedom for the students to decide upon certain aspects. According to Priemer (2011), this openness can refer to different dimensions which are not completely separable from each other:

- *Openness in terms of content* which an inquiry can be assigned to (e.g. Millar, Tiberghien & Le Maréchal, 2002). The openness gradually varies between a pre-defined content and the subsequent question that should be investigated, to an intermediate state with a selection of pre-defined contents from which students can choose, to complete freedom in the choice of topic under investigation.
- *Openness in terms of strategy* of the inquiry (e.g. Millar et al., 2002). This dimension refers to the approach with which a specific investigation is planned to be carried out. This could include the decision on qualitative or quantitative approaches and the development of a design. Again, the openness can be gradually varied from specific instructions, to sole indications and hints on possible approaches, and to complete freedom without scaffolding.
- *Openness in terms of methods used* in the inquiry (e.g. Fischer & Draxler, 2001). This dimension refers to the choice of methods used to carry out the above-mentioned strategy. This includes the choice of instruments and materials. The openness obviously varies from specific guidelines on what methods to choose, to a pre-defined selection of methods from which students can choose, to complete freedom without scaffolding.
- *Openness in terms of the number of possible solutions* (e.g. Blömeke, Risse, Müller, Eichler & Schulz, 2006). This dimension refers to the variety of different solutions that may result from an inquiry. The openness varies from one single solution (e.g. in an investigation on the interdependencies between current, voltage and resistance) to many possible solutions (e.g. in an investigation on the question how the melting point of a substance can be changed).
- *Openness in terms of the number of different solution processes* (e.g. Blömeke et al., 2006). This dimension refers to the variety of different solution processes that are possible in order to get some result (e.g. find out the material which an object consists of which could be determined through measuring buoyancy, measuring the resistivity, or electrolysis). The openness varies from one single solution process to many possible solution processes.

A distinct feature of inquiry in science education is its openness of the student activities. This openness may concern different dimensions such as the content or the strategies applied. Openness in terms of content, for example, means that the content of an inquiry is not pre-defined by the teacher, by a school book or by some other external player but by the student who engages with the inquiry him/herself.

For the purpose of this study, the dimensions of openness as defined in Priemer (2011) are taken as a basic conceptualisation. The two reasons for this decision are that these dimensions are concrete enough to be used as categories for coding classroom units and that it is the most wide-spread and prominent conceptualisation in the German-speaking community.

3.1.4 The student-oriented nature of inquiry-based education – a transition to the subsequent sub-chapters on formative assessment

The openness of student activities in the context of inquiry as described in the above section has consequences of the roles of the student and the teacher. These will be outlined here.

Hinrichsen and Jarrett (1999) explain the learning process of students from a constructivist perspective: „Students need to personally construct their own understanding by posing their own questions, designing and conducting investigations, and analysing and communicating their findings. Students need to have opportunities to progress from concrete to abstract ideas, rethink their hypotheses, and adapt and retry their investigations and problem-solving efforts” (Hinrichsen & Jarrett, 1999, p. 5). According to this understanding, the learning of students is facilitated by “the opportunity to undertake ‘research activities’ instead of just carrying out routine ‘cook-book experiments’” in problem-based or inquiry teaching approaches (European Commission, 2004, p. 125). The European Commission also states that there should be an emphasis on combining minds-on and hands-on activities, for example by the use of open-ended tasks, by combining different activities, and by self-directed learning. Similarly, the EU-project S-TEAM suggests that students should engage in authentic and problem-based learning activities with more than one correct answer; in experiments and hands-on activities, including searching for information; in self-regulated learning sequences where the student autonomy is emphasized; and in discursive argumentation and communication with peers (Jorde, Olsen Moberg, Rönnebeck & Stadler, 2012).

Apparently, such activities lead to a shift in the roles of students and teachers compared to more traditional approaches to teaching and learning (Kessler & Galvan, 2007). Whereas students take the active part of engaging in scientifically oriented questions, in developing explanation from evidence, in considering alternative explanations, and in communicating and justifying their explanations (Euler, 2011), teachers “lead students to develop the skills necessary for inquiry and the understanding of science concepts through their own activity and reasoning” (McLoughlin, Finlayson & van Kampen, 2012, pp. 14–15). So the teacher role is the initialization and coaching of the inquiry process (Kessler & Galvan, 2007). One approach to supporting students in their inquiry activities is formal formative assessment as outlined in the following sub-chapters.

A distinct feature of inquiry in science education is its student-oriented nature. This means that in inquiry, the students should be the main actors rather than the teachers. This has the consequence that the role of the teacher is being a coach. One approach to coaching students in their inquiry learning is formative assessment which will be introduced in the subsequent sub-chapters.

3.2 The concept of formative assessment

Assessment can be described from different perspectives (e.g. European Commission, 2004, p. 137): “(1) traditionally, as the function of evaluating student achievement for grading and tracking, (2) as an instrument for diagnosis to give students and teachers continual feedback about learning outcomes and difficulties, and (3) as a means to enable broader knowledge about the conditions behind and influences on students’ understanding and competence (e.g. in international large-scale assessments)”.

At the level of classroom practice, the first and the second perspective, summative and formative assessment, are relevant. In both cases, data about student learning is collected and interpreted (see Figure 3). The difference lies in the purpose of that data collection (e.g. Harlen, 2013): In the case of summative assessment, the data is used for summarizing and reporting about student performance at a particular time and, for this reason, it is also called ‘assessment of learning’ (ARG, 2002). In the case of formative assessment, the data is collected in order to decide about next steps in learning. Therefore, formative assessment is also called ‘assessment for learning’ (ARG, 2002).

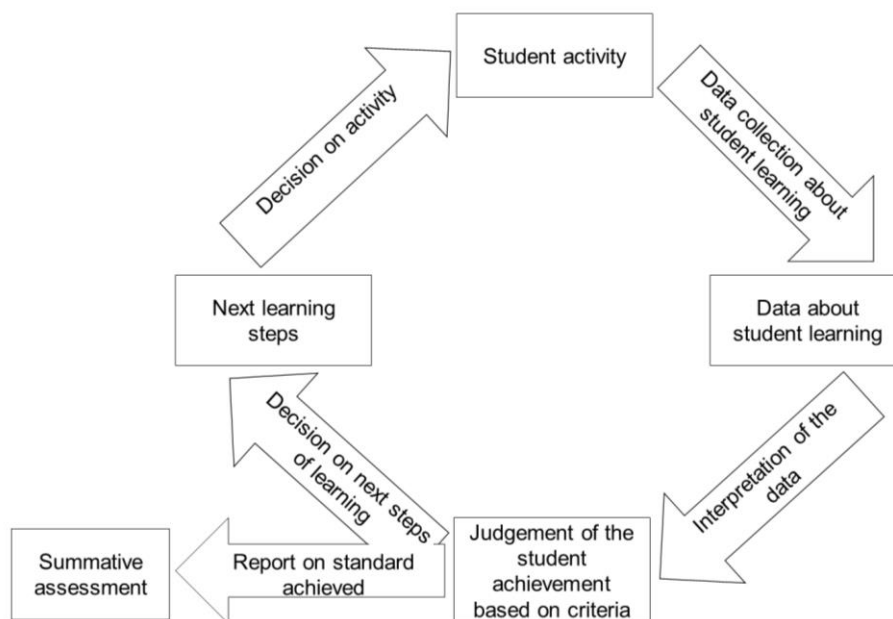


Figure 3: Formative and summative assessment (based on Harlen, 2013).

In this sub-chapter, theoretical perspectives on formative assessment will be introduced first (section 3.2.1). That first section will serve as a theoretical background. In the subsequent sections, the general steps which a formative assessment activity consists of will be introduced (section 3.2.2), and afterwards, two systems to characterize formative assessment activities will be introduced (degree of formality in section 3.2.3 and cycle lengths in section 3.2.4).

The aim of this sub-chapter is to introduce models and definitions which are tangible enough to afterwards deduce categories for the empirical part of the study from them. These categories will be used to describe the formative assessment activities of the teachers’ trials which are investigated as part of the research questions.

3.2.1 Terminology

Formative assessment has the purpose of assisting learning and for that reason is also called ‘assessment for learning’. It involves processes of “seeking and interpreting evidence for use by learners and their teachers to decide where the learners are in their learning and where they need to go and how best to get there” (ARG, 2002, p. 2). The term was first used in the context of evaluation and assessment by Scriven (1967) and Bloom (1969) and further refined by Ramaprasad (1983), Popham (2008), and Sadler (1989). Formative assessment is much more than the “long neglected bridesmaid in the testing party” (Cizek, 2010, p.4):

As the definition from the Assessment Reform Group ARG above implies, formative assessment “refers to the collaborative processes engaged by educators and students for the purpose of understanding the students’ learning and conceptual organisation, identification of strengths, diagnosis of weaknesses, areas for improvement, and as a source of information that teachers can use in instructional planning and students can use in deepening their understandings and improving their achievement” (Cizek, 2010, p.6). So both the teacher and the student can potentially use the information gathered (Stiggins, 2005). In teacher-centred units, formative assessment may help the teacher to plan subsequent steps in teaching. In the context of student-oriented activities, feedback to the student has the multiple purpose of “enhancing desired skills, refining learning of valuable objectives, and fostering intrinsic motivation” (Cizek, 2010, p. 7).

Apart from elaborating on a definition, a number of authors have attempted to describe formative assessment by key characteristics. As the author of one of these compilations says, however, not all of these characteristics need to be fulfilled in order to consider an activity formative (Cizek, 2010). The first key characteristic is that formative assessment is an integral part of teaching and learning (e.g. Bell & Cowie, 2001; Birenbaum et al., 2006). The underlying understanding is that learning and assessment are not considered separate but integral. The second key characteristic is that formative assessment requires students to take responsibility of their own learning and to self-monitor their progress towards the learning goals (Cizek, 2010). This idea shows the proximity between formative assessment and self-regulated learning which will be described in more detail in sub-chapter 3.3.4. Thirdly, formative assessment activities are planned but still allow teachers for some adjustment and flexibility in order to meet individual student needs (OECD, 2005a). Fourthly, formative assessment is continuous, informs students of their current level of achievement and provides feedback and guidance to learners on how to improve their learning by scaffolding information and focusing on the learning process (Looney, 2011; Wilson & Sloane, 2000). The fifth key characteristic concerns the feedback: Feedback is specific, is given in a timely manner, and is linked to specific criteria (Sadler, 1989) which are clearly specified in advance and represent valuable educational outcomes (Looney, 2011).

Formative assessment is assessment for learning. The steps involved in such activities, on a more practical level, will be introduced in the next section.

3.2.2 Four steps for formative assessment

There are a number of methods and activities to formatively assess student learning. However, in general, the process follows four steps: Articulation of expectations, diagnosis, feedback, and use of this feedback (e.g. Andrade & Valtcheva, 2009; Gregory, Cameron & Davies, 2000; Paris & Paris, 2001; Ross, Hogaboam-Gray & Rolheiser, 2002; Stallings & Tascione, 1996). These steps will be introduced in more detail here.

The articulation and sharing of expectations allows the student to show and the teacher to diagnose the student level of achievement. It can therefore be seen as the “starting point for effective formative assessment” (Ruiz-Primo et al., 2010, p. 140). This articulation may take place in different formats: The criteria could be explicitly communicated by the teacher (e.g. Andrade & Valtcheva, 2009); elaborated together with the students in the classroom (e.g. Black, Harrison, Lee, Marshall & Wiliam, 2004); or the criteria could also be implicitly clear within a particular unit (e.g. if they are the same for a period longer than this particular unit).

The diagnosis of the student level of achievement can take place based on different types of data such as written reports; informal conversation with the students; presentations; and many others. Formal formative assessment should be implemented at “junctures or waypoints” (Ruiz-Primo et al., 2010, p. 142) where the teacher might wish to check whether his/her students progress in their learning as expected before they move on to the next phase. Once the information on the students’ learning is gathered, the teacher needs to analyse and interpret it in order to understand where the student is compared to the overarching goals.

The feedback to the learner to plan subsequent actions: Based on the analysis and interpretation mentioned in the last paragraph, a decision about the next steps in learning has to be taken. In the context of student-oriented activities, this could be feedback to the students on how to proceed. The decision on the use of feedback depends on the degree of planning involved, the level of formality of the feedback, the type of data on student learning searched for, and the type of feedback (Furtak & Ruiz-Primo, 2008).

The use of the feedback could, in the context of student-oriented activities, either consist of a revision of the original artefact based on the feedback or of a transfer of the feedback to a similar, new situation (e.g. Andrade & Valtcheva, 2009; Paris & Paris, 2001).

In practice, formative assessment typically consists of four steps: Articulation of expectations, diagnosis, feedback, and use of this feedback.

For the purpose of this study, these four steps are taken as the criteria to decide whether an interaction in the classroom is formative assessment or not. To be classified as formative assessment, an interaction must include the articulation of expectations (in the form of criteria provided by the teacher; elaboration of criteria with students; or implicitly); diagnosis (based on data such as written artefacts; observations; oral data); provision of feedback to students (on the results of the diagnosis); and use of this feedback (revision of original artefact or transfer to similar situation).

3.2.3 Degree of formality in formative assessment

Formative assessment activities can be described by their amount of formality (Shavelson et al., 2008): Depending on the amount of planning involved, the nature and quality of the data sought, and the nature of the feedback given to students by the teacher, it ranges from on-the-fly (totally informal) to an intermediate level (planned-for-interaction; formal) and ends totally formally embedded in the curriculum. A similar distinction is made by Cowie & Bell (1999), who articulate two kinds of formative assessment processes. The first one is 'interactive' where no specific activity is undertaken and the assessment simply arises from the learning activity. The second kind of formative assessment processes is 'planned' where activities are undertaken that specifically allow for formative assessment.

One way of classifying the many methods of formative assessment is to order them by their degree of formality. A second classification will be introduced in the next section.

For the purpose of this study, the categories from Shavelson et al. (2008) are used as a basic conceptualisation. The two reasons for this decision are that 1) the categories cover the varieties of formative assessment methods employed in Swiss teaching practice and 2) that the labelling appeared more appropriate than in the similar model from Cowie & Bell (1999). The labelling of the later would imply that only on-the-fly assessment is interactive.

However, for reasons concerning the research methods, only planned-for-interaction formative assessment is investigated in this study.

3.2.4 Cycle lengths in formative assessment

A second possibility to classify formative assessment activities is the length of a cycle. The length of a cycle refers to the time between the formative assessment activity itself and the use of the information or feedback derived from it. Wiliam (2010) develops three types of cycles that characterize formative assessment (see Table 1):

Table 1: Cycle lengths for formative assessment (from Wiliam, 2010).

Type	Focus	Length
Short-cycle	Within and between lessons	Minute by minute: 5 seconds to 2 hours Day by day: 24 to 48 hours
Medium-cycle	Within and between instructional units	1 to 4 weeks
Long-cycle	Across marking periods, quarters, semesters	4 weeks to 1 year

One way of classifying formative assessment activities is by the length of an assessment cycle. The term refers to the time span between which a student product is assessed and the opportunity to employ the feedback received on that product.

3.3 Mechanisms in formative assessment that support learning

In their meta-study, Black and William (1998) showed that formative assessment methods produce significant learning gains and hereby confirmed earlier reviews by Natriello (1987); Crooks (1988); and Kluger and DeNisi (1996). The effects are among the largest ever identified for educational interventions (Hattie, 2009; Looney, 2011). Black and Wiliam (1998) gathered 250 international studies focussing on the use and impact of formative assessment at different school levels. From these, they selected the 40 studies that were conducted under ecologically valid circumstances (controlled experiments conducted in the students' usual classroom setting and with their usual teacher). The formative assessment activities included effective feedback; questioning; comprehensive approaches to teaching and learning featuring formative assessment; and student self- and peer-assessment. The meta-study also revealed that formative assessment has particularly high effects for low-achieving students and that there is not only an impact on achievement but also on motivation. One of the emerging questions, though, is through what mechanisms student learning is supported.

In this sub-chapter, three approaches explaining how formative assessment supports student learning will be introduced. The approaches are: Sharing assessment criteria (section 3.3.1), feedback (section 3.3.2), and self-regulated learning (section 3.3.3).

The aim of this sub-chapter is to provide some theoretical background on the effects of formative assessment on student learning.

3.3.1 Sharing assessment criteria as part of formative assessment

Sharing the intentions of learning (e.g. of a lesson or a unit) and identifying clear assessment criteria is a central part of formative assessment (e.g. Black, Harrison, Lee, Marshall & Wiliam, 2003; Mansell, James & the Assessment Reform Group, 2009). Assessment criteria are guidelines under which work will be assessed (Goodrich, 1996). Such criteria can be established in different ways (Panadero & Alonso-Tapia, 2013): They can be externally set by the teacher; they can be formulated by negotiation between the teacher and the students; or they can be set internally by the individual student.

Since assessment criteria clarify the achievement goals and the "features of excellent performance" (Shepard, 2000, p. 11), they help the students in shaping their learning (Andrade & Valtcheva, 2009; Boekaerts & Corno, 2005). Furthermore, the clear achievement goals support students in choosing their learning strategies and therefore contribute to an enhanced self-regulation (Looney, Laneve & Moscato, 2005; Shepard, 2000).

The common understanding of assessment criteria between students and teachers is one approach to explain the effects of formative assessment on student learning.

3.3.2 Feedback as part of formative assessment

Feedback is "information provided by an agent (such as a teacher, peer, book, parent, self) regarding aspects of one's performance or understanding" (Hattie & Timperley, 2007, p. 81). It is considered a powerful metaphor that underlies the theory of formative assessment in the international literature (Wiliam, 2010). The idea of feedback originates from the field of system engineering and basically states that the information collected in a system must have some effect on this very system (Ramaprasad, 1983). The same author concludes that "feedback is the information about the gap between the actual level and the reference level of a system parameter which is used to alter the gap in some way" (Ramaprasad, 1983, p. 4). This conclusion already implies that the gap may be altered in different ways. For example "through a number of different cognitive processes, including restructuring understandings, confirming to students that they are correct or incorrect, indicating that more information is available or needed, pointing to directions students could pursue, and/or indicating alternative strategies to understand particular information" (Hattie & Timperley, 2007, p. 82). Focussing on these cognitive processes, Hattie & Timperley (2007) suggest in their feedback model that feedback should provide the students with information concerning three major questions:

“Where am I going? (What are the goals?) How am I going? (What progress is being made toward the goal?) Where to go next? (What activities need to be undertaken to make better progress?” (Hattie & Timperley, 2007, p. 86).

A number of studies have investigated the effect of feedback practices on student learning (e.g. Allal & Lopez, 2005; Hattie & Timperley, 2007; Köller, 2005; Shute, 2008). The studies are hard to compare; however, common themes can be recognized (Wiliam, 2010). One of these themes that emerge from various studies is that effective feedback should look forward, not back. The main concern should be ‘what next’, not ‘what was right and what did I get wrong’ (Wiliam, 2010). Another perspective is that different types of feedback may have different levels of effectiveness on different learning aims (Dempster, 1991; 1992). Shute (2008), for example, reports that procedural learning may be more effectively supported by immediate feedback whereas delayed feedback seems to be more efficient for higher-order skills (Shute, 2008).

An attempt to structure these different types of feedback has been undertaken by Hattie & Timperley (2007). They define four levels to which feedback can be targeted: “Task level (how well tasks are understood/ performed), process level (the main processes needed to understand /perform tasks); self-regulation-level (self-monitoring, directing, and regulating of actions); and self-level (personal evaluations and affect, usually positive, about the learner)” (Hattie & Timperley, 2007, p. 87). In the context of formative assessment of competences as laid out in the preceding sections, the process level is of major importance. It will therefore be focussed on in the next few sentences. Feedback about the processing of the task is “more specific to the processes underlying tasks or relating and extending tasks” (Hattie & Timperley, 2007, p. 93). It implies a “deep understanding of learning <that> involves the construction of meaning (understanding) and relates more to the relationships, cognitive processes, and transference to other more difficult or untried tasks” (Hattie & Timperley, 2007, p. 93). The two authors just cited mention two major groups of process-level feedback: Feedback that enhances the students’ strategies to detect errors and feedback that acts as cueing mechanisms.

However, feedback mechanisms do not follow a simple input – output model: Kulhavy (1977) stresses that feedback does not automatically lead to the desired behaviour from the recipient. Instead, feedback can be modified or rejected rather than accepted by the recipient. Furthermore, feedback is not always provided consciously by a teacher, peer, or parent, but can also be sought by the recipient or even detected by the recipient without being intentionally sought (Hattie & Timperley, 2007).

Feedback for students is an approach to explaining the effects of formative assessment on student learning. It is the approach which is mentioned most prominently in the literature.

3.3.3 Formative assessment and self-regulated learning

As already mentioned, feedback is considered the pivotal part of formative assessment by many authors, particularly in the sources written in English language. However, it has been criticised (Bose & Rengel, 2009; Clark, 2012; Nicol & Macfarlane - Dick, 2006; Panadero & Alonso-Tapia, 2013; Perrenoud, 1998) that this framework is too narrow: “[...] part of the feedback given to pupils in class is like so many bottles thrown out to sea. No one can be sure that the message they contain will one day find a receiver [...]” (Perrenoud, 1998, p. 87). The same author therefore suggests that formative assessment should not be identified by its practice (the presence of feedback) but by its effect: The contribution to the regulation of learning processes (Perrenoud, 1991; 1998).

Self-regulation is “an active, constructive process whereby learners set goals for their learning and then attempt to monitor, regulate, and control their cognition [...]” (Pintrich & Zusho, 2002, p. 250). Pintrich (2000) established four processes how self-regulated learning is supported by formative assessment: Firstly, the students become active participants in their learning process through formative assessment. This is particularly evident in peer- and self-assessment. Secondly, the students get the opportunity to “monitor, control, and regulate certain aspects of their own cognition, motivation, and behaviour as well as

some features of their environment” (Pintrich, 2000, p. 454) in formative assessment activities. The third process is the provision of criteria or standards which enable students to compare their own learning to and to decide whether their learning process should be adapted. Fourthly, “self-regulatory activities mediate a three-way dynamic between personal and contextual characteristics and performance” (Pintrich, 2000, p. 453). Apparently, some of the aspects mentioned in Pintrich (2000) relate back to earlier parts this theory chapter and cannot be separated from those completely, such as the sharing of assessment criteria as part of the formative assessment (see 3.3.1). But the powerful aspect of this section is probably the understanding that formative assessment helps students to become independent learners.

Enhancement in self-regulated learning is an approach to explaining the effects of formative assessment on student learning.

3.4 Methods of formative assessment for the context of inquiry-based science education

Many formative assessment methods have been described for use in science learning (e.g. Angelo & Cross, 1993; Keeley, 2008). Some of them are best usable for diagnosing the students' conceptual understanding (e.g. the often mentioned traffic lights where students indicate their understanding with a colour code; Keeley, 2008). Others provide scaffolds in autonomous learning activities such as inquiry-based learning (Baron & Darling-Hammond, 2008; Black & Harrison, 2004; Ruiz-Primo & Furtak, 2007). Amongst these are more informal, spontaneous methods like on-the-fly and more formal methods. A number of the later methods for formal formative assessment will be discussed in the following sections.

In this sub-chapter, three formal methods of formative assessment that are suitable for inquiry-based learning will be introduced. The methods are: Written teacher assessment (section 3.4.1), peer-assessment (section 3.4.2), and self-assessment (section 3.4.3).

The aim of this sub-chapter is to introduce the formative assessment methods and the research that has been conducted on them as a background for the empirical part of the study: These are the three methods that were trialled by the teachers in the study and on which respective results will be presented.

3.4.1 Written teacher assessment

Written teacher assessment and feedback can be provided in the form of instructional rubrics (Andrade, 2005; Arter & McTighe, 2001; Burke, 2006; Moskal, 2003) and open comments (e.g. Black & Harrison, 2004). In many cases, the two varieties are combined so that the teachers use an assessment template that has both an assessment rubric (which includes the learning goals) for formative use and open space for comments on it.

Open comments show, similar to rubrics, individual problems and specific strengths of a piece of work (Black et al., 2003). However, they are more directed to making improvements towards the learning goals and to show concrete next steps whereas the instructional rubrics were more focussed on communicating the current state of achievement (Nunes, 2004; Santos & Dias, 2006; Stracke & Kumar, 2010).

Potentials and challenges of the assessment method

Empirical evidence (Panadero & Jonsson, 2013) suggests that written teacher assessment can improve student learning by five mechanisms: By increasing transparency in articulation the goals, by reducing student anxiety in clarifying the expectations, by aiding the feedback process, by improving the student self-efficacy, and by supporting student self-regulation.

However, a number of difficulties have also been identified: Emphasis is put on the prerequisites that only with clear, measurable goals available to the students and an activity that is suitable for assessing those criteria, written teacher assessment can be used meaningfully (Jonsson, 2014; Luft, 1999; Moskal, 2003). Furthermore, the content and use of written teacher assessment has to be explained to students in order to make it a useful tool for learning (Andrade & Du, 2005; Moni & Moni, 2008).

Considering instructional rubrics, both So and Lee (2011) and Bharuthram (2015) find that the rubrics are used in a rather unconscious manner and without exploiting the full range of possibilities and potential of the tool. The detriments of the use of rubrics from the perspective of teachers and lecturers, as mentioned in a number of studies, include issues of time and the difficulty to formulate the criteria in a rubric in a way that is understandable to students (e.g. Bharuthram, 2015; Luft, 1999).

Regarding open comments, four difficulties have been identified in the literature. The first difficulty is the criteria the written teacher assessment focusses on. A number of studies in language learning at different school levels show that teachers' written assessment often targets superficial aspects (spelling, grammar) instead of global issues such as the development of competences (Connors & Lunsford, 1993; Hargreaves & McCallum, 1998; Leki, 2006; Schwartz, 1984; Stern & Solomon, 2006). Also for science education, Bruno

and Santos (2010) report teacher challenges in how to select what to comment on and how to avoid giving away part of the answer. Weaver (2006) adds that open teacher comments are not always related to goals or assessment criteria.

Secondly, most comments appear as non-comments (grades, numeric scores, or symbols), or as evaluative comments indicating whether the student is right or wrong (Ruiz-Primo & Li, 2013). As a result of this, the percentage of descriptive comments providing information on how and why the work is right or wrong and prescriptive comments offering additional information on what to do next is low (Ferguson, 2009; Glover & Brown, 2006; Hernández, 2012; Walker, 2009).

The third difficulty is the students' understanding of the feedback received. Glover & Brown (2006) show that students have trouble understanding teachers' feedback because of the language it is formulated.

Lastly, the effectivity of written comments also seems to depend on the student commitment: Hyland claims that the use of written comments by individual students seems to vary due to "individual differences in needs and student approaches to writing" (Hyland, 1998, p. 255).

For the purpose of this study, written teacher assessment as a formative assessment method is defined as diagnosis and associated feedback provided in a written form by the teacher.

3.4.2 Peer-assessment

Terminology

Peer-assessment follows the idea of "activating students as instructional resources for one another" (Leahy, Lyon, Thompson & Wiliam, 2005, p. 21): Students take both the role of the assessor and the assessee by assessing each other's work. The aim of peer-assessment is to assist peers in identifying the strengths and weakness or their work and to provide suggestions for improving it (Dochy, Segers & Sluijsmans, 1999; Topping, 2003). Peer-assessment can be based on rating instruments or checklists which may be designed by others beforehand or developed by the user group themselves (Falchikov, 1991).

Potential and challenges of the assessment method

A number of advantages and challenges that are associated with peer-assessment have been identified in the literature. The advantages of peer-assessment are, firstly, that feedback from peers who had the same difficulties in the learning progress might suggest direct ways to overcome those difficulties, and formulate them in a language that is naturally used by the students (Black et al., 2004). Secondly, students who assess their peers' work engage in cognitively demanding activities, such as critical thinking (Hanrahan & Isaacs, 2001; Harlen, 2007; Lin, Liu & Yuan, 2001; Lindsay & Clarke, 2001; Topping, 2003; Tsivitanidou, Zacharia & Hovardas, 2011). Thirdly, students get the opportunity to see examples of other students' work. This can potentially lead to self-assessment: By comparing their own work to that of their peers, students can be prompted to reflect on their own learning achievements (Hanrahan & Isaacs, 2001; Lin et al., 2001; Topping, 1998; 2010). Fourthly, peer-assessment may be easier to accept since it is perceived less authoritative than feedback from adults and therefore open to negotiation (Cole, 1991; Topping, 2010). Fifthly, feedback from peers can be more immediate, timely, and individualized than feedback from the teacher (Topping, 2010) simply because there are many more students than teachers in a classroom. Lastly, providing feedback to peers develops the social, communicative, metacognitive and other personal and professional skills on the way (Topping, 2010).

The challenges of peer-assessment identified in the literature are the following: When doing peer-assessment, students need to judge the performance of a peer. This needs a certain degree of knowledge in the field that is assessed (Topping, Smith, Swanson & Elliot, 2000). Furthermore, students need to communicate the judgments to their peers and to provide constructive feedback about their learning process. This needs communication skills (Black et al., 2003). Thirdly, the recipients need to critically review the feedback and decide on the actions to be taken: Since peer-feedback might include flaws, the recipients need to filter it

and then decide whether there is a need to adopt the peers' suggestions and to revise their work (Sluijsmans, 2002). Fourthly, peer-assessment costs lesson time for organisation, training and monitoring, particularly in the beginning, if it should be provided at a good level of quality (Topping, 2010). Lastly, social processes influence and contaminate the validity and reliability of assessment provided by peers (Topping, 2010).

For the purpose of this study, peer-assessment as a formative assessment method is defined as diagnosis and provision of feedback conducted between peer-students.

3.4.3 Self-assessment

Terminology

"Self-assessment is a process of formative assessment during which students reflect on the quality of their work, judge the degree to which it reflects explicitly stated goals or criteria, and revise accordingly" (Andrade, 2010, p. 91). Similarly to peer-assessment, self-assessment prompts the students to think about the quality of their own work themselves instead of having the teacher as the source of judgements (Andrade, 2010). Contrary to self-evaluation and self-grading, self-assessment is "done on drafts of works in progress in order to inform revision and improvement: It is not a matter of having students determining their own grades." (Andrade & Valtcheva, 2009). Self-assessment is also distinguished from reflections: Self-assessment is task-specific whereas reflections are more general judgements about strong and weak abilities (Harrington, 1995).

Potentials and challenges of the assessment method

Goodrich (1996) investigated the conditions under which criteria-referenced self-assessment can be beneficial: These prerequisites include the awareness of the value of self-assessment; access to clear criteria on which to base the assessment; a specific task or performance to assess; models of self-assessment; direct instruction in and assistance with self-assessment; practice; cues regarding when it is appropriate to self-assess; and opportunities to revise and improve the task or performance. A number of authors suggest different tools to support self-assessment such as rubrics (Arter & McTighe, 2001; Moskal, 2003; Burke, 2006; Smit & Birri, 2014), examples of good practice (Black et al., 2003; Hanrahan & Isaacs, 2001) or traffic lights (Black & Harrison, 2004).

Three basic clusters of benefits are reported in the literature: Firstly, self-assessment boosts learning and achievement. Research on the effects of student self-assessment covers a wide range of areas including writing (Andrade, Du & Wang, 2008; Evans, 2001; Hart, 1999; Wilcox, 1997; Yancey, 1998), mathematics (Adams, 1998; Ross et al., 2002; Stallings & Tascione, 1996), and science (Duffrin, Dawes, Hanson, Miyazaki & Wolfskill, 1998; White & Frederiksen, 1998). Secondly, self-assessment is central to self-regulation because students must be aware of the goals of the task and checking their progress towards them (Nicol & Macfarlane-Dick, 2006; Schunk, 2003; Zimmermann & Schunk, 2004). This helps the students to become reflective practitioners, that means that they are able to reflect critically upon their own professional practice (Dochy & Moerkerke, 1997; Falchikow & Boud, 1989). Lastly, self-assessment helps to monitor and formally account one's own learning (Boud, 1990; Harvey & Knight, 1996; Kwan & Leung, 1996; Pintrich, 2000; Zimmerman & Schunk, 2004).

Both teachers' and students' opinions on self-assessment have been investigated. Students believe that it helps them to keep focussed on key elements of assignments, to learn the respective content, to improve their ability to identify weaknesses and strengths in their work, to increase their motivation and decrease anxiety (Andrade & Valtcheva, 2009). Another study on students' opinions about self-assessment brought similar results but additionally showed that students are convinced that self-assessment has a positive effect on their grades and on the quality of their work (Andrade & Du, 2007). Teachers find self-assessment rewarding and practicable (Black & Harrison, 2001).

Hanrahan and Isaacs (2001) investigated the challenges associated with self-assessment as perceived by students and teachers: The students in their sample found that the process was time-consuming and not in all cases worth the effort. Furthermore, the students reported that the self-assessment was not always taken serious since it does not result in a grade. From the perspective of the teachers, the following challenges are mentioned (Hanrahan & Isaacs, 2001): Firstly, it is difficult to provide clear assessment criteria. Secondly, students need time to practice self-assessment and to become familiar with the respective instruments; and thirdly, the timing can be a challenge.

For the purpose of this study, self-assessment as a formative assessment method is defined as reflections of the students focussing on the quality of their work and appropriate measures to reach the goals or criteria.

3.5 Teacher concepts of and self-efficacy in formative assessment

Conceptions include people's beliefs, attitudes and intentions (Brown, 2008; Thompson, 1992). They have a large effect on how people behave (Ajzen, 2005). More specifically, the relation between beliefs teachers have and actions they take have been shown in a number of studies (Pajares, 1992; Rubie-Davies, Flint & McDonald, 2011; Woolfolk Hoy, Davies & Pape, 2006). In the context of assessment, it has been found that the teachers' conceptions of teaching, learning and curricula influence how they teach and what students achieve (Calderhead, 1996; Clark & Peterson, 1986; Pajares, 1992; Thompson, 1992). However, that influence does not need to go so far as to a "neat correspondences between teachers' beliefs and their practices" (Marshall & Drummond, 2006, p. 144).

In this sub-chapter, two constructs that relate to conceptions will be introduced in the subsequent two sections: Teacher concepts of assessment and feedback (section 3.5.1) and teacher formative assessment efficacy (section 3.5.2).

The aim of this study is to explore possibilities and challenges in the implementation of formative assessment methods in daily teaching. This does not only require a change in the teachers' practices but also in the teachers' beliefs and attitudes. The aim of this sub-chapter is introduce selected aspects of these concepts in more detail to provide the background for the respective results in the empirical part of the study.

3.5.1 Teacher concepts of assessment

Brown (2004) distinguishes four basic concepts of what teachers think assessment is. The first concept is that assessment has the purpose of improving teaching and learning in the classroom – the intentions of formative assessment (Black & Wiliam, 1998). The second concept is that the purpose of assessment is to control the quality of a teacher or a school (Firestone, Mayrowetz & Fairman, 1998). The third concept is that assessment checks the learning of the individual students (Dixon, 1999; Hill, 2000) which reflects the intentions of summative assessment. The final concept is that assessment has "no legitimate place within teaching and learning" (Brown, 2004, p. 305).

Brown, Harris and Harnett (2012) focus on New Zealand teachers' (N=518) concepts of feedback and their relation with formative assessment: They constructed a survey with items on the purposes, the types, the validity and the timing of feedback mainly drawing from Hattie and Timperley (2007) on the four types of feedback (see section 3.3.2) and from Irving, Harris and Peterson (2011) on different purposes of feedback (irrelevance; improvement; reporting and compliance, encouragement). They find that the teachers "endorsed feedback factors [...] of using assessment and feedback to improve learning" (Brown et al., 2012, p. 974). The teachers stressed the idea of "involving students in generating and using feedback to improve their work and develop autonomy [...]" (Brown et al., 2012, p. 974). Furthermore, the authors found that "encouragement and protection of student self-esteem were considered as only minor aspects of these learning-oriented conceptions of feedback" (Brown et al., 2012, p. 975). So the effect of feedback on student learning seems to be more relevant than the effect on student motivation. The authors argue that this result is plausible since "almost by definition, teachers are interested in improving the learning of children and adolescents and it is expected that teachers would endorse a learning-orientation in their feedback" (Brown et al., 2012, p. 976). Another result of the study is that the conceptions between primary and secondary teachers were largely identical (Brown et al., 2012).

A study on the relation between teachers' beliefs and their practice revealed that "teachers [...] progress and change in how they relate their values to their practices within a project such as LHTL <Learning how to learn>" (Marshall & Drummond, 2006, p. 144). They conclude that teachers' beliefs on assessment are not stable but may be influenced by the teachers' collaboration in projects.

Four different teacher concepts on assessment have been reported in the research literature, one of them relating closely to the aims of formative assessment. Furthermore, the teachers' concepts of feedback have been investigated. The main aim of feedback was considered to be an enhanced student learning. These

conceptions have been reported to be largely identical across school levels. Another study showed that teachers' beliefs on assessment are not stable but can change in collaborative projects.

3.5.2 Teacher formative assessment efficacy

Self-efficacy is a construct developed by Bandura (1977; 1982). The term is defined as "beliefs in one's ability to organise and execute actions required to handle future situations. Put more simply, self-efficacy refers to a person's confidence that they can do what they have to do" (Brigido, Borrachero, Bermejo & Mellado, 2013, p. 3). Self-efficacy has two dimensions (Bandura, 1982): Personal self-efficacy which is described as "self-assessment of one's ability to perform the task" (Mintzes, Marcum, Messerschmidt-Yates & Mark, 2013, p. 1202) and outcome expectancy which is described as "his or her expectation that performing the task will result in a desirable outcome" (Mintzes et al., 2013, p. 1202). Behaviour is, according to Bandura (1977; 1982), based on both factors.

Self-efficacy is contextually dependent (Bandura, 1997; Shell, Colvin & Brunning, 1995). For the context of science teaching, a number of studies investigated the relations between self-efficacy and teaching and revealed that a number of teaching behaviours such as risk taking and the trial of innovative ideas (Ashton & Webb, 1986) but also the extent of inquiry and other student-oriented teaching methods (Bhattacharyya, Volk & Lumpe, 2009; Czerniak, 1990) are related to self-efficacy in science teaching.

The most widely used instrument for measuring self-efficacy in science teaching was developed by Enochs and Riggs (1990). It describes self-efficacy with two scales: personal self-efficacy belief and outcome expectancy. In the context of formative assessment, the personal self-efficacy belief could be reflected in a teacher's confidence about implementing a new formative assessment method in his or her teaching. The outcome expectancy, on the other hand, might be a judgement about how likely it is that such a method, if implemented well, will support the students in their learning. However, no empirical studies focussing on teacher self-efficacy in formative assessment have been found. A model suggesting how teachers' inquiry teaching skills and formative assessment could boost each other has been developed, though (Dolin & Evans, 2013). It proposes that both spontaneous and structured formative assessment will lead to a good student learning outcome. This positive outcome will enhance both the student's and the teachers' self-efficacy. The higher teacher self-efficacy is expected to promote the use of more inquiry-based science education. So the different factors are expected to interact in a positive cycle.

Teacher self-efficacy can be changed (Ashton & Webb, 1986; Ramey-Gassert & Shroyer, 1986). Bandura (1977; 1982) recognized four contributors to self-efficacy: The first contributor is mastery experience. In the context of formative assessment, it could be a successful implementation of a formative assessment method. Secondly, vicarious experiences are mentioned. They could, in the context of formative assessment, consist of the observation of a peer-teacher who formatively assesses in his or her classroom. The third contributor is social persuasion, such as the emotional support from a community of like-minded professionals. Fourthly, physical and emotional factors are mentioned, such as the perception and interpretation of signs of stress.

Pre-service programs as well as in-service professional development programs have been reported to increase teachers' self-efficacy in science teaching (Cone, 2009; Hechter, 2011; Palmer, 2006; Yoon et al., 2006). No studies focussing specifically on formative assessment in science education have been found. A number of studies, however, focus on a particular type of professional development programs, the so called 'professional learning community' (PLC) which is a process "in which the teachers in a school and its administrators continuously seek and share learning, and act on their learning" (Hord, 1997, p. 6). This particular type of professional development programs was found to have several positive effects on teacher knowledge, beliefs and attitudes on the one hand, but also on teacher instructional practice on the other hand which were all related to teacher self-efficacy (Fulton & Britton, 2010).

Self-efficacy is a construct that describes a persons' belief in her / his ability to perform a particular task. The construct consists of two dimensions: The personal self-efficacy belief (self-assessment of the person's performance) and the outcome expectancy (expectations that the outcome of the persons' performance will be good). Self-efficacy is context-dependent; this means that it is different for different tasks and circumstances. Teacher self-efficacy in various domains is reported to be changeable, for example through pre-service programs as well as through in-service professional development and particularly through so-called professional learning communities in which teachers and school administrators collaborate in improving a particular aspect of their teaching.

In the context of science teaching, the most widely used instrument for measuring self-efficacy was developed by Enochs and Riggs (1990). For the purpose of this study, the instrument was adapted to focus on formative assessment rather than on science teaching generally.

3.6 Obstacles to putting formative assessment into practice and measures of support

Several OECD publications stress the importance of formative assessment in European educational systems which are, however, dominated by summative assessment practices (Looney, 2011; OECD, 2005a). In order to change this, a number of authors identify barriers and suggest measures of support.

The barriers that prevent formative assessment activities from becoming part of regular teaching practice are located at different levels: At the level of educational systems and therefore pointing towards political and administrative stakeholders but also at the level of the individual classroom, so pointing at the teachers. Some of the barriers cannot be clearly placed on one of these levels, however.

In this sub-chapter, six challenges identified from the literature will be introduced: The relation between formative and summative assessment (section 3.6.1); and the link between different groups of stakeholders (section 3.6.2) which are both located at the level of educational systems; the perceptions of formative assessment (section 3.6.3) which relate to both the level of educational systems and individual classrooms; the abilities of teachers (section 3.6.4) and logistics (section 3.6.5) which are both located mainly at the level of the individual classroom; and assessment of competences (section 3.6.6) which both cannot be placed on one of the above-mentioned levels.

The aim of this study is to explore possibilities and challenges in the implementation of formative assessment methods in daily teaching. A part of the empirical data will focus on the challenges and potential measures of support as perceived by the teachers who collaborated in the study. The aim of this sub-chapter is introduce problems and measures of support identified in the literature to serve as a background for the empirical data of this study.

3.6.1 Relation between formative and summative assessment

One cluster of challenges refers to the role of formative assessment in the assessment system. Both Looney (2011) and Pedder (2006) perceive a tension between formative and summative assessment whereas Gitomer and Duschl (2007) prefer to speak about a lack of coherence. The European Commission does not only identify this challenge in the classrooms but also includes international assessments like PISA and TIMSS in the system: “Although the results <of large-scale international assessments like PISA and TIMSS> may be used to identify strengths and weaknesses in each country, there is a danger that these studies may trivialize the purpose of schooling by its implicit definition of how educational 'quality' might be understood, defined and measured. It is likely that national school authorities put undue emphasis on these comparative studies, and that curricula, teaching and assessment will be 'PISA-driven' in the years to come” (European Commission, 2004, p. ix).

Many authors therefore suggest measures to utilise possible synergies in promoting learning and to improve the continuity at least between formative and summative assessment at a classroom level, or even to integrate the international large-scale assessments. It is suggested that continuity could be developed through the alignment of both formative and summative assessment with the curriculum goals underlying teaching and learning in the classroom (OECD, 2005a; 2013; Pellegrino & Hilton, 2012; Shavelson et al., 2008; Wilson & Sloane, 2000). Summative assessment is one of the main driving forces in education and a key character to any educational system, signaling priorities for curricula and instruction (Binkley et al., 2012; Gardner, Harlen, Hayward, Stobart & Montgomery, 2010; Harlen, 2007). If the alignment between curriculum on the one hand and formative and summative assessment on the other hand was clearly perceived by students, the impact on their goal setting could be very strong (Allal, 2010; Chudowsky & Pellegrino, 2003).

Furthermore, it is suggested that continuity could be developed through high-quality feedback on the learning outcomes for students: So the use of summative data for formative purposes. This is expected to help regulating the students' subsequent efforts in learning (Allal, 2010; Williams & Ryan, 2000). Adding to this,

Looney (2011) and Watson (2006) wonder whether performance data from international assessments could be constructed in a way that they provided relevant information for individual teachers, too.

A third suggestion is to develop continuity within the assessment framework is by ensuring that standards of validity, reliability, feasibility, and equity are met in both formative and summative assessment (American Association for the Advancement of Science, 1998).

Fourthly, some authors suggest that continuity could be developed through student involvement in summative assessment. This form of assessment inevitably needs a judgment formulated by a professional (teacher, examiner, or other expert) about the quality of student learning. It is possible, nevertheless, to develop some degree of active student engagement. In a portfolio assessment used for summative purposes, for example, students could participate in the selection of the work samples to be included and could write self-reflective commentaries that accompany their work (Allal, 2010; Black et al., 2004).

One of the barriers that prevent the use of more formative assessment in classroom as identified in the literature is the relation between formative and summative assessment. It is suggested that the two facets should be better aligned; that synergies should be used more efficiently; and that the role of the students and the teachers in assessment should be strengthened.

3.6.2 Link between different groups of stakeholders

This cluster of challenges addresses the lack of coherence in terms of aims and purposes of assessment between the policy, school and classroom level (Looney, 2011). Some of the measures of support that were suggested above might also apply here. Furthermore, OECD (2005a) and Shavelson et al. (2008) claim that stronger links between policy, practice and research should be built and institutionalised in order to foster collaboration between practitioners, curriculum experts, and assessment experts.

One of the barriers that prevent the use of more formative assessment in classroom as identified in the literature is that the links between policy, school and classroom level are not strong enough and should therefore be institutionalised.

3.6.3 Perceptions of formative assessment

The third cluster of challenges emerging from the literature refers to perceptions of and beliefs about formative assessment. Formative assessment is feared to be too resource-intensive and time consuming to be practical (Looney, 2011) but also to be 'soft', non-quantifiable and therefore not important by policy makers, administrators, teachers and other stakeholders (Looney, 2011). Many of the suggestions that were mentioned for the first cluster of challenges may also apply here. Furthermore, an enhanced accountability of formative assessment methods so that teachers feel confident about their acceptance by school administrations but also by the public is suggested by the American Association for the Advancement of Science (1998).

One of the barriers that prevent the use of more formative assessment in classrooms as identified in the literature is the perception of various groups of stakeholders that formative assessment is not valid and not important. Many of the suggestions that were brought up on the relation between formative and summative assessment also apply here.

3.6.4 Teacher assessment literacy

This cluster of challenges emerging from the literature concentrates on the abilities of teachers. A lack of formative assessment skills is reported (Bennett, 2011; Stiggins, 1999) but also a lack of reasonably deep understanding of the contents taught (Bennett, 2011). Regarding the formative assessment skills, a number of specific problems are addressed: The teachers' difficulties in identifying a clear focus of a formative assessment activity such as a relevant learning goal (Cizek, 2010) but also the teachers' difficulties in appro-

privately accommodating the formative assessment activities so that both students and the teacher can optimally benefit from it (Cizek, 2010). Furthermore, teachers are reported to have troubles in developing and administering good formative assessment activities, e.g. developing good questions to probe student learning (Swaffield, 1998; Yin et al., 2008) but also interpreting student responses or in formulating next steps for instruction and providing specific feedback (Herman, Osmundson & Silver, 2010; Yin et al., 2008). Both Cizek (2009) and Popham (2008) find that teachers have difficulties in assessing objectively and being aware of the classroom assessment bias. Finally, the challenges are not limited to the capabilities of the teachers but also cover their motivation and beliefs concerning the importance of formative assessment (Ruiz-Primo et al., 2010) which relates back to section 3.6.3.

Many authors claim teacher professional development programs to enhance the teachers' and the school leaders' assessment literacy: Their skills to integrate tools for formative assessment in their classroom activities; their skills in terms of diagnosis of students' competences and in terms of providing support to students; an awareness of the different factors that may influence the validity and reliability of results; the capacity to make sense of data, to identify appropriate actions and to track processes (Alkharusi, 2011; American Federation of Teachers, National Council on Measurement in Education, & National Education Association, 1990; Brookhart, 2011; Looney, 2011; OECD, 2005a; Pedder, 2006; Schwartz & Allal, 2000; Wiliam, 2006). The teachers should also be introduced to ways of integrating assessment and instruction and be supported in understanding their role as coaches rather than as transmitters of knowledge (American Association for the Advancement of Science, 1998).

Apart from the provision of professional development, a second approach to tackle teachers' assessment literacy emerged from the literature: In the field of education, the teachers' understanding and acceptance of innovation is crucial for success (Wilson & Sloane, 2000). It is therefore suggested in the literature that the frame for teachers to innovate should be provided (OECD, 2005a): For example to review their assessment questions and discuss them with peers (Ayala et al., 2008; Black & Wiliam, 1998), but also to exchange experiences in order to train diagnosis- and feedback skills (Schwartz & Allal, 2000).

One of the barriers that prevent the use of more formative assessment in classroom as identified in the literature is the teachers' assessment literacy. It should be enhanced through professional development and through the provision of platforms for teachers to innovate.

3.6.5 Logistics

This cluster summarizes logistic, practical challenges: Large classes, extensive curriculum requirements, and the difficulty of meeting diverse and challenging student needs (OECD, 2005a; Looney, 2011) take a lot of the teachers' time and energy. They therefore are not able to take the efforts of developing formative assessment activities (Bennett, 2011; Cizek, 2010). As means of support to overcome these barriers, the provision of tools for formative assessment is suggested (e.g. American Association for the Advancement of Science, 1998). Furthermore, the possibilities of technology should be taken advantage of more rigorously (Chudowsky & Pellegrino, 2003; Looney, 2011).

One of the barriers that prevent the use of more formative assessment in classroom as identified in the literature is logistic aspects such as large classes which prevent teachers from developing their own formative assessment techniques. It is therefore suggested that such instruments should be provided.

3.6.6 Assessment of competences

The last cluster of challenges affects both formative and summative assessment: It is the modelling of higher-order skills. Watson (2006) writes about inquiry and development of competences in mathematics education and wonders if non-linear pathways of the students' development can be described at all. Following this, the same author questions "how [...] such descriptions <could> be used by teachers and students without reducing [...] enquiry to a rubric without purpose?" (Watson, 2006, p. 301). In this context, several authors (e.g. Looney, 2011) suggest that the many gaps in research and development should be addressed.

One of the barriers that prevent the use of more formative assessment in classrooms as identified in the literature is that assessment of competences per se is not easy. More research is needed to approach the problem.

3.7 Effects of programs implementing formative assessment

A number of empirical papers address attempts to implementing formative assessment in teaching and learning practice. According to Maier (2015), this is, however, a difficult issue. As examples, the author cites a number of German studies which have found that assessment for formative purposes is rarely used in Germany (Engel, 2008; Grotlüschen & Bonna, 2008; Maier, 2011). This is explained the traditionally high importance of summative assessment (Breidenstein, Meier & Zaborowski, 2012; Köller, 2005; Zaborowski, Meier & Breidenstein, 2011).

Estimates from other countries are similar; e.g. for the US where particularly formal (e.g. written, planned-for-interaction) formative assessment activities are rarely and unsystematically used in regular classroom practice (Morrison & Lederman, 2003). Yin et al. (2008) conclude that “simply embedding assessments in curriculum does not guarantee improved learning and teaching. Teachers need tremendous support using assessment in their teaching practice. Moreover, teachers must also figure out how best to adapt formative assessment to their needs and the need of their students” (Yin et al., 2008, p. 356).

In the international literature, a number of projects address this issue from different perspectives and these will be introduced in this sub-chapter: The implementation of formative assessment through professional development which can be seen as top-down-approach (section 3.7.1); and bottom-up approaches where teachers develop their own formative assessment strategies (section 3.7.2).

The aim of this study is to explore possibilities and challenges in the implementation of formative assessment methods in daily teaching from data generated in the collaboration with twenty teachers from two school levels. This sub-chapter will help to situate the design of the study (the way of collaboration) as introduced in chapter 5 in a broader context.

3.7.1 Professional development of teachers

A number of studies have focussed on the effect of professional development on the quality of teachers' formative assessment abilities (Brookhart, Moss & Long, 2010; Mertler, 2009; Sato, Wei & Darling-Hammond, 2008). These studies show that the teachers' abilities in formative assessment can be developed (Maier, 2015) even though “transferring new teaching approaches into practice is not straightforward” (Hondrich, Hertel, Adl-Amini & Klieme, 2015, p. 1). Several authors add that implementing formative assessment activities in their teaching is a challenge for teachers (e.g. Black & Atkin, 1996; Furtak et al., 2008; Smith & Gorard, 2005; Tierney, 2006).

Results of the professional development programs of the “National Board Certification” in the US

Sato et al. (2008) investigated the effects of the professional development programs of the “National Board Certification” in the US on the formative assessment practices of mathematics and science teachers. The quality of the teachers' assessment practices was measured in six dimensions: Usage of the assessments; variability, quality and coherence of the assessments; clarity of the learning aims; opportunities for self-assessment; adaption of the teaching based on the assessment; quality and appropriate fit of the feedback to the learners. The research group found significant improvements in the quality of the formative assessment practices compared to a control group which had not received the professional development, particularly in the two dimensions ‘variability of assessment methods used’ and ‘usage of student data for supporting student learning’.

Results of Co²CA in Germany

The Co²CA project in Germany aimed at supporting primary school teachers in implementing formative assessment in their daily science teaching practice (Bürgermeister et al., 2011). The professional development program focussed firstly on the basics of formative assessment, secondly on subject-specific factors, and thirdly on the provision of helpful feedback to students. The researchers investigated the teachers' implementation fidelity (Dusenbury, Brannigan, Falco & Hansen, 2003) which describes the extent to which the teachers' classroom practice reflected the professional development program (Hondrich et al., 2015). The

results showed that the teachers were highly capable in direct application of the formative assessment strategies introduced: A nearly perfect implementation frequency and a high quality of the feedback were found. The transparency of the formative process, e.g. the explication of the aim of the formative assessment and feedback was not as well enacted (Hondrich et al., 2015). The authors of the study concluded that “primary school teachers are able to implement most aspects of a curriculum-embedded formative assessment intervention when it is combined with supportive materials and professional development workshops – even given a challenging subject like science” (Hondrich et al., 2015, p. 15).

However, the teachers from the same intervention group had difficulties in transferring the formative assessment strategies to a new topic. They used the assessment activities at a lower frequency, with a lower transparency of enactment and provided feedback of a lower quality. The authors concluded that “devising the necessary materials proved too difficult or time-consuming for teachers to keep up with the intended rate of four assessments within two weeks” (Hondrich et al., 2015, p. 15). In consistence with earlier studies (e.g. Desimone, 2009; Gresham, 1989), the study shows the importance of explicit training and provision of supportive materials in the implementation of formative assessment. The authors also mention that a promising approach is to provide a platform for teachers to develop their own tools. This collaboration with peer teachers and appropriate support would take a lot of time but could enhance the flexible use of tools across different topics (Postholm, 2012; Wiliam, Lee, Harrison & Black, 2004). This idea will be elaborated on in more detail in the subsequent section 3.7.2.

The first type of studies on the implementation of formative assessment introduced is by professional development of teachers. The two studies described in detail report positive effects of the professional development programmes on the teachers’ formative assessment practices. The first study reports a higher variability of assessment methods used and better usage of student data for supporting learning. The second study reported very good direct application of the strategies provided in the professional development program but problems with the transfer to other contents and topics.

3.7.2 Teachers developing their assessment

The idea of teachers developing their own formative assessment strategies and practices was first brought up in Black and Wiliam (1998): “[...] the improvement of formative assessment cannot be a simple matter. There is no ‘quick fix’ that can be asses to existing practice with promise of rapid reward. On the contrary, if the substantial rewards of which the evidence holds out promise are to be secured <the positive effects of formative assessment on student achievement>, this will only come about if each teacher finds his or her own ways of incorporating the lessons and ideas that are set out above into her or his own patterns of classroom work. This can only happen relatively slowly, and through sustained programmes of professional development and support.” (Black & Wiliam, 1998, p. 15). Reasoning for this claim is provided in Wiliam et al. (2004): The “difficulty of ‘putting research into practice’ is not the fault of the teacher. But nor is it a failing in the research. Because our understanding of the theoretical principles underlying successful classroom action is weak, research can never tell teachers what to do. Indeed, given the complexity of classrooms, it seems likely that the positivist dream of an effective theory of teacher action – which would spell out the ‘best’ course of action given certain conditions – is not just difficult and a long way off, but impossible in principle” (Wiliam et al., 2004, p. 51).

Results of a collaborative project between researchers and teachers in England

Following the approach outlined above, a project with 1 ½ years of collaboration between researchers with teaching experience and 24 teachers from six selected schools from England was set up (Black et al., 2003; Wiliam et al., 2004). The teachers were not instructed on the exact procedures of assessing and teaching. Instead, the general principles of formative assessment were introduced and discussed in workshops. These principles included rich questioning, comment-only marking, sharing criteria with learners, and student peer-assessment and self-assessment (Wiliam et al., 2004). Based on the introductions, the teachers decided individually what principles they wished to concentrate on and developed their own practical implementations of formative assessment activities. The teachers were supported in the planning and realising their

formative assessment throughout a school year: Firstly in a series of in-service sessions (six-and-a-half day in total) where all teachers came together; and secondly in regular visits to schools where the teachers were observed by project staff. These visits were also used for in-depth-discussions.

Afterwards, the students' achievement in an externally mandated test was measured. In the frame of a 'local design' (Wiliam et al., 2004), the achievement of the students who had attended lessons with formative assessment was compared to other students. 'Local design' means that the control groups were set up depending on the local conditions: They could consist of parallel classes taught by the same teacher in the previous year, or parallel classes taught by peer teachers. The majority of effect sizes were around 0.2 to 0.3; the mean was calculated as 0.34 (Black et al., 2003; Wiliam et al., 2004). The authors of the study concluded that the teacher professional development program helped teachers to develop and sustainably integrate formative assessment activities in their teaching.

On the other hand, a high variability in the quality and the sustainability of the implemented activities was noted (Black et al., 2003; Wiliam et al., 2004). This is reflected in the different ways in which the participating teachers adopted formative assessment strategies: Black et al. (2003) distinguish between four teacher groups, namely the trailers which are characterised as "teachers who had attempted strategies but had not embedded any strategies into their practice" (Black et al., 2003, p. 28); static pioneers who are described as "teachers who were successful with one or two key strategies and who had restricted themselves to these" (Black et al., 2003, p. 28); moving pioneers who are "teachers who were successful with one or two key strategies, but having routinized these were looking for other ways to augment their practice" (Black et al., 2003, p. 28); and finally the experts who have "formative assessment strategies embedded in and integrated in practice" (Black et al., 2003, p. 28).

Results of a school development project in Switzerland

Assessment can be considered a typical focus in school development projects (Maier, 2015). An example from Switzerland will be presented here. Smit (2009) researched on a school development project which aimed at changing the assessment culture at lower secondary school level in the Kanton Zug/Switzerland. The main focusses of the project included: Stronger orientation on learning aims in instruction and in assessment; implementing student self-assessment; discussions on the students' levels of achievement between teachers; and enhanced selection processes for the passover from one school year to the next. As part of the empirical study, the students evaluated the quality of the formative assessment in their classes. A positive correlation was found between formative assessment activities and individual self-assessment of the students' levels of achievement. So the formative assessment appeared to have an influence on the students' self-confidence. The relation was particularly clear for female student with a negative self-assessment. Formative assessment activities appeared to compensate for such low self-esteem. However, the students' engagement with the formative assessment they received varied within classes.

The qualitative results of the above-mentioned study (Smit, 2009), revealed that the development of new, innovative formats of assessment is easier when several teachers collaborate and exchange ideas. This can be facilitated by subject group meetings or other networks. Another aspect that appears to enhance the uptake of formative assessment is an established practice of student-oriented teaching with a certain degree of openness in the tasks. This seems to allow for more flexible planning of formative assessment activities.

The second type of studies on the implementation of formative assessment introduced is teachers developing their own assessment practices alone, collaboratively, or with the support of researchers. Positive effects on student achievement and in the students' self-confidence in science, particularly amongst female students, are reported in the two studies introduced here. At the same time, a high variability in the quality and the sustainability of the activities developed and implemented is also reported.

3.8 Inquiry and assessment in Switzerland

Compared to other countries, Switzerland has a highly de-centralised educational system. Firstly, there is a strong federal tradition and the individual *Kantone* (states) have a considerable degree of freedom in issues concerning education. Secondly, the teachers of all school levels have a high autonomy in their teaching and far-reaching responsibilities which are connected to this autonomy. These responsibilities include, for example, the design of almost all summative tests and the respective decisions on grades.

So the development of comprehensive curricula for the compulsory school levels, valid for all *Kantone* (states) in each of the linguistic regions in the last years is considered a big step towards harmonisation of the compulsory school levels. In these comprehensive curricula, educational standards for different school years have been defined in several subjects including science. At the level of *Gymnasium* (upper secondary school), harmonisation started somewhat earlier: A curriculum, the *Rahmenlehrplan nach MAR*; (EDK, 1994) valid for the *Gymnasien* of all linguistic regions has been valid since 1994, in a revised version since 2007.

This sub-chapter will provide an introduction to the educational system in Switzerland where the study described took place. The focus will be on inquiry-based education (section 3.8.1) and on assessment (sections 3.8.2; 3.8.3).

The aim of this sub-chapter is to provide the contextual background of this study: It will help to situate the empirical results in the national context.

3.8.1 Inquiry-based education

The above-mentioned curriculum for the compulsory school levels and its underlying competence model contain many elements of inquiry-based science education, as will be introduced below. To a smaller extent, this is also true for the curricula at the *Gymnasium* level. But little is known about the extent to which these curricular guidelines are put into practice. However, the empirical research on the practice of science education in Switzerland that relates to inquiry will be outlined towards the end of the section.

Competence model as a basis for the curricula at the compulsory school levels

As laid out above, new, competence-oriented curricula which are valid for all *Kantone* (states) of a linguistic region are being implemented at the moment for the compulsory school levels. For the German-speaking regions of Switzerland, the curriculum is named '*Lehrplan 21*' and has started to be implemented in the different *Kantone* (states) since 2015/16.

The basis for the curricula at the compulsory school levels in the science subjects is a competence model and minimal standards which were developed in the HarmoS project (HarmoS: Konsortium HarmoS Naturwissenschaften, 2008; EDK, 2011). The Swiss competence model for science consists of three axes: Skills (*'Handlungsaspekte'*), contextual domains, and achievement levels (Labudde, Nidegger, Adamina & Gingsins, 2007; Labudde, 2007). So the Swiss model explicitly distinguishes between skills and competences (Labudde et al., 2007; Labudde, 2007): Examples of skills in the model include, amongst others, 'to ask questions and to investigate' and 'to exploit different sources of information' (HarmoS: Konsortium HarmoS Naturwissenschaften, 2008; EDK, 2011). Examples of contextual domains in the model include 'motion', 'force', 'energy', 'structure and changes of matter', and others (HarmoS: Konsortium HarmoS Naturwissenschaften, 2008; EDK, 2011). The model can be considered an operationalisation of Weinert's concept of competences (Weinert, 2001) which is predominant in the German-speaking countries. Weinert's concept conveys the understanding that a competence is the ability (skill) to do something in a particular context (domain). The science skills from HarmoS are well aligned with the processes of inquiry as explained in section 3.1.1 (see Table 2) and it can be expected that many of the skills from HarmoS can be fostered with inquiry-based education.

Inquiry in the science curriculum for the compulsory school levels

Following the elaboration of the competence model and the decision to define minimal standards for compulsory school, curricula for the different subjects and the different linguistic regions were developed. The science curriculum in the German-speaking part of Switzerland defines minimal standards for grades 2, 6, and 9 (D-EDK, 2014), where grade six is the end of primary school, whereas grade 9 represents the end of lower secondary school. Apart from defining these basic standards based on the competence model introduced in the above paragraph, curriculum 21 also contains the extended list of skills from the competence model: The ‘*Denk-, Arbeits- und Handlungsweisen*’ (manners of thinking, working and acting; D-EDK, 2014). The introduction of these skills can be perceived as an additional signal to promote inquiry-based education.

Table 2 displays the alignment of inquiry activities as defined in Bell et al. (2010) with the science competence model and the manners of thinking, working and acting. It can be seen that apart from ‘prediction’ as a last step in the processes of inquiry from Bell et al. (2010), all processes are reflected in both the competence model and in the curriculum for the compulsory school levels.

Table 2: Alignment of inquiry activities as defined in Bell et al. (2010), the science competence model (HarmoS, 2008) and the manners of thinking, working and acting (D-EDK, 2014) from the curriculum of the compulsory school levels.

Main processes of inquiry learning Bell et al. (2010), see sub-chapter 3.1	Competence model for compulsory school science (HarmoS, 2008; EDK, 2011) <i>Grundkompetenzen</i>		Curriculum 21 (D-EDK, 2014) <i>Lehrplan 21</i>
	Skill ‘To ask questions and to investigate’ <i>Handlungsaspekt ‘Fragen und untersuchen’</i>	Skill ‘to communicate and to exchange’ <i>Handlungsaspekt ‘Mitteilen und austauschen’</i>	Manners of thinking, working and acting <i>Denk- Arbeits- und Handlungsweisen</i>
Orienting and asking questions	To pose questions, problems and hypotheses <i>FU2: Fragen, Probleme und Hypothesen aufwerfen</i>		To question <i>Fragen</i>
Hypothesis generation			To assume <i>Vermuten</i>
Planning	To choose and use suitable tools, instruments and materials <i>FU3: Geeignete Werkzeuge, Instrumente und Materialien auswählen und verwenden</i>		To investigate <i>Untersuchen</i> To experiment <i>Experimentieren</i>
Investigation	To conduct explorations, investigations or experiments <i>FU4: Erkundigungen, Untersuchungen oder Experimente durchführen</i>		
Analysis and interpretation			
Model			
Conclusion and evaluation	To reflect upon results and methods <i>FU5: Über Ergebnisse und Untersuchungsmethoden nachdenken</i>		
Communication		To describe, present and reason <i>MA1: Beschreiben, präsentieren und begründen</i>	To document <i>Dokumentieren</i> To communicate <i>Mitteilen</i> To exchange <i>Austauschen</i>
Prediction			

Inquiry in the curricula of the science subjects at upper secondary school (Gymnasium)

At the level of *Gymnasium* (upper secondary school), harmonisation started somewhat earlier: A curriculum, the *Rahmenlehrplan nach MAR*, (EDK, 1994) for all *Gymnasien* of all linguistic regions has been valid since 1994, in a revised version since 2007. This *Rahmenlehrplan* does not formulate competences but distinguishes between contents, skills and attitudes that should be covered before the final *Matura* exam for every subject. The *Rahmenlehrplan* serves as a basis for the more detailed syllabus that is elaborated by each *Gymnasium* or, depending on the region, by all *Gymnasien* of a *Kanton* (state). The alignment of the skills with the processes of inquiry after Bell et al. (2010) can be seen in Table 3. Apparently, the skills in the *Rahmenlehrplan* are formulated more generally than the inquiry processes in Bell et al. (2010) so that one skill typically covers several processes. Also, the *Gymnasium* curricula cover fewer inquiry processes than the curriculum for the compulsory school levels does. The situation varies between the three subjects, though: A considerable portion of the inquiry processes are covered in the physics- and in the biology curriculum, whereas the chemistry curriculum does not contain inquiry skills as defined in Bell et al, 2010.

Table 3: Alignment of inquiry activities as defined in Bell et al. (2010) and the curriculum for the *Gymnasium* (*RLP nach MAR*).

Main processes of inquiry learning Bell et al. (2010), see sub-chapter 3.1	Curriculum for <i>Gymnasium</i> level (EDK, 1994) <i>Rahmenlehrplan nach MAR</i>		
	Skills physics (EDK, 1994, p. 108) <i>Grundfertigkeiten Physik</i>	Skills chemistry (EDK, 1994, p. 111) <i>Grundfertigkeiten Chemie</i>	Skills biology (EDK, 1994, p. 115) <i>Grundfertigkeiten Biologie</i>
Orienting and asking questions			To discover, observe and document situations and processes <i>Entdecken, Beobachten und Dokumentieren von Zuständen und Prozessen</i>
Hypothesis generation			To develop working hypotheses <i>Arbeits-hypothesen entwickeln</i>
Planning	To plan, set up, conduct, analyse and interpret simple experiments <i>Einfache Experimente planen, aufbauen, durchführen, auswerten und interpretieren</i>		To plan and conduct meaningful experiments with living species responsibly, to record and represent [data] in words and graphically, to critically test, evaluate and conclude <i>Sinnvolle Experimente mit lebenden Organismen verantwortungsvoll planen und durchführen, protokollieren, sprachlich und graphisch darstellen, Aussagen kritisch prüfen und werten, sich ein Urteil bilden und Methodenkritik üben</i>
Investigation			
Analysis and interpretation			
Model	To develop models and apply them on specific situations <i>Modelle gewinnen und auf konkrete Situationen anwenden</i>		
Conclusion and evaluation			
Communication	To observe and describe phenomena and technical processes in own words, formulate physical relations mathematically but also in colloquially <i>Naturabläufe und technische Vorgänge beobachten und mit eigenen Worten beschreiben, physikalische Zusammenhänge mathematisch, aber auch umgangssprachlich formulieren</i>		
Prediction			

Research on science teaching practice at primary school level

Little research has been conducted on the practice of science teaching at different school levels in Switzerland. One of the few studies took place in the context of the elaboration of the science standards for the compulsory school levels in 2006 (HarmoS: Konsortium HarMoS Naturwissenschaften, 2008). Consequently, it focussed on the teaching practice at primary and lower secondary school level. 362 primary school teachers and 37 lower secondary school teachers from 5 *Kantone* (states) were asked about the conditions (time structures; infrastructure), about the topics they covered in their science teaching, and about the forms of teaching and learning in a questionnaire. Selected results from the primary school teachers will be provided below. These teachers were all involved in the pre-service training of student teachers as *Praktikumslehrpersonen* which means that they regularly mentor student teachers in their practical training.

The results show that the primary school teachers typically (at least 75% positive answers; more in some *Kantone*) had the possibility to organise time slots of more than two lessons for their science units, if needed. This opportunity was typically used 1-3 times per school year. Most teachers (percentage not indicated in HarMoS: Konsortium HarMoS Naturwissenschaften, 2008) perceived the room situation to be problematic, with the teachers not having enough rooms for their science teaching. All teachers had the opportunity to visit close-to-nature environments, but the variation in the use of this opportunity was high between teachers. The use of school books was also investigated: It varied substantially between regions and also between grades. The majority of teachers (percentage not indicated in HarMoS: Konsortium HarMoS Naturwissenschaften, 2008) reported that they had only a small collection of materials and tools for their science teaching and specifically for experimenting in the classroom. The teachers were also asked about the forms of teaching and learning they employ. Around 10% of the teachers indicated that they often did open units whereas at least 50% of the teachers (more than 70% at grades 3/4 and 5/6) indicated that they never did open units. Between 20-35% (depending on the grade) of the teachers answered that they often did inquiry teaching, and a smaller portion of teachers answered that they never did that. Traditional teaching and guided workshop activities (*Werkstattunterricht, Postenarbeit*) were employed more often than inquiry teaching at primary school level. When asked about the scientific practices employed, the teachers considered observing, exploring and investigating as well as searching for information the most common practices at their school level.

Research on the science teaching practices at lower secondary school level

The PhD study of Johannes Börlin (Börlin, 2012; Börlin & Labudde, 2014) investigated the enactment of experiments in physics education at lower secondary school in Germany, Switzerland and Finland. He analysed 99 videotapes of double lessons (2*45min) on the connection between electrical energy and power that involved an experimentation phase. A third of these videotapes were recorded in Swiss classrooms. Börlin found that the time dedicated to experiments within these double lessons varied significantly between the countries, with an average of 42 mins (out of 90mins) for the Swiss cases. The time dedicated to experiments in the Swiss cases was mostly used for conducting the experiments (46%), to a smaller extent to the post-processing (30%) and the preparation (20%). Both qualitative and quantitative experiments were conducted in the Swiss cases, with no significant difference in the frequency. The experiments were mostly used either as demonstration experiments conducted by the teacher or as student experiments in groups of three or more students. For the conduction, lab equipment (such as power supply, volt meter) was used rather than everyday objects.

Deep structure analyses did not reveal big differences between the three countries investigated. In the preparation and in the post-processing phases of the experiments, relevant aspects of physics content were covered. However, these aspects were not sufficiently linked to the aims and the procedures of the experiments. Instead, the relations were only shown on a general level. Research questions and hypotheses were not clarified and the experiments were divided into small steps which were conducted one after the other. Therefore, both students and teachers tended to lose the overview of the whole unit. Reflexions on neither the results nor on the process of the experiments did occur frequently.

Typically, the rationale behind the experiments was either to conduct measurements or to illustrate a physical law (qualitatively or quantitatively). Both functions relate to a low or medium cognitive demand (Börlin & Labudde, 2014). Less than 20% of all experiments covered one of the following aims: Solving or realising a technical problem; explore a phenomenon; visualise a physical concept; getting familiar with a measurement device; or enhance the understanding of scientific methods.

Research on the science teaching practices at the upper secondary school level

EVAMAR I and EVAMAR II, two large surveys at *Gymnasium* (upper secondary school) level as well as MUPET (e.g. Dreyer, 2015), a smaller study, are too general to draw any conclusion on inquiry or assessment in science education. Instead, their focus is on the students' ability to study at university; on their achievement in different subjects; and on their satisfaction with their education at the *Gymnasium*.

So just two empirical studies researching inquiry practice or related topics at upper secondary science teaching were found: Labudde (2000) analysed the degree to which the principles of constructivism are embedded in physics education at the *Gymnasium*. Constructivism was conceptualised as a construct containing four dimensions: The individual dimension; the dimension of content; the social-communicative dimension; and the dimension of teaching and learning methods (*unterrichtsmethodische Dimension*). He analysed a number of different data sources: Official documents concerning the physics curriculum at the *Gymnasium* level; students' answers (152 classes; 671 students) on a questionnaire that was added to the 1995 TIMSS survey; and structured individual interviews with physics teachers. Labudde found that the constructivist approach is clearly supported in all documents analysed (legal regulations; *Maturitätsanerkennungsreglement*, *Rahmenlehrplan*; a number of position papers). The documents explicitly mention and demand for numerous elements of constructivism in the individual dimension (pre-conceptions; self-regulation); in the dimension of content (links to everyday life; links to humans; openness of problems etc.); in the social-communicative dimension (discourse, role of the teacher etc.); and in the dimension of educational methods (experiments conducted by teachers and students; project work etc.).

The analysis of the students' answers in Labudde's study on how they perceive their physics education showed mixed results: On the one hand, the students confirmed that their prior knowledge and pre-conceptions were taken into account in the physics classes; that demonstration experiments conducted by the teacher were included in many of their physics lessons; and that other indicators of constructivist teaching were present. On the other hand, the students also said that only few relations to humans, to the society, and to the scientific community were revealed; that only few experiments could be conducted by the students themselves; that project work was hardly ever enacted; and that the main teaching methods were teacher-centred question-and-answer sessions (*fragend-entwickelnder Unterricht*); experiments conducted by teachers, and frontal inputs by the teachers.

Labudde concluded that despite the positive aspects that emerged from the analysis, the physics education did generally not follow a constructivist approach as demanded in the *Maturitätsanerkennungsreglement* and *Rahmenlehrplan*. To some extent, the situation may have improved with the introduction of the *Maturaarbeiten* (matura thesis) and the introduction of *Integrationsfächer* (interdisciplinary subjects) in the last decade.

Widmer Märki (2011) investigated the possibilities of conducting and assessing interdisciplinary science units at the *Gymnasium* (upper secondary school level). She collaborated with 27 teachers. Her data included individual interviews, teaching plans and materials of interdisciplinary units. Widmer Märki found that in the interdisciplinary units analysed, assessment without grades played a minor role. However, a number of summative assessment strategies were used: The classic written summative test; presentations and reports; and oral exams. Furthermore, more innovative forms of assessment were also trialled: Grading of posters, concept maps, portfolios, as well as instruments for assessing the working process. Some teachers also included the students' self-assessments in their grading. Overall, Widmer Märki concluded that embedding more than one assessment strategy in a single interdisciplinary unit allowed for the assessment of several competences.

Widmer Märki stressed the importance of selecting data on student learning from which the desired competence can be appropriately diagnosed. Taking joined-up thinking as an example, Widmer Märki suggested that these joins should be visually (e.g. in mind maps) or verbally represented (e.g. in oral presentations). These forms of assessment have the advantage that the students have the opportunity to present their level of performance rather than the teacher constructing suitable test questions. Another aspect discussed by Widmer Märki was that innovative forms of teaching is not only challenging for the teachers but also for the students. In order not to overburden teachers and students, Widmer Märki suggested that innovative forms of assessment should be combined with more traditional forms in a balanced manner.

At the compulsory school levels, new curricula valid for a whole linguistic region each have started to be implemented. The basis for these curricula in the science subjects is a competence model with minimal standards developed in the HarmoS project. Both the competence model and the science curricula foster competence-oriented teaching-learning approaches that are rooted in moderate constructivism, such as inquiry.

At upper secondary school level (*Gymnasium*), the harmonisation of the curricula started earlier (the presently used frame for the curricula has been valid since 1997). At this school level, it is distinguished between knowledge, skills and attitudes in the curricula. The need for inquiry-based teaching might, to some extent, be indicated in the skills domain.

In a small number of studies, the actual situations in the science classrooms have been explored at primary, lower and upper secondary school levels. The studies are small-scale and cannot provide a valid overview of the science teaching practice in Switzerland in general. However, the picture that emerges is that there are indications for inquiry-based and constructivist teaching at all school levels but that there is room for improvement in various aspects.

3.8.2 Summative assessment

Summative assessment tradition

Switzerland has no culture of high-stake large-scale assessment at any school level. Instead, the individual teacher is typically responsible for the design of tests and the grading. As mentioned earlier, this can be interpreted as a sign of the high teacher autonomy in the country.

At the compulsory school levels, a small number of regional large-scale assessments (regional because of the federal tradition) have been established in the last fifteen years. The main purpose in most cases of these assessments is to survey the educational system but not to yield a direct impact on individual students.

At the level of the *Gymnasium*, discussions to coordinate summative assessments in subject groups within individual schools or even between different schools (e.g. the question if or not the final exam at the end of year 12 should be the same within a *Gymnasium* or even within a *Kanton*) become more and more prominent.

Literature related to summative assessment practices in Switzerland

Grades are generally accepted and thought to be necessary in Switzerland, particularly among parents and students (e.g. Dzelili, 2009). So efforts to reduce the importance of traditional tests and grading are difficult to communicate and their political acceptance is not easily achieved (Fischer, 2009), even though typical flaws of summative assessments are well-known among teachers and researchers (e.g. Frey & Frey-Eiling, 2004).

When it comes to assessment practices in the individual classroom, Rothenbacher (2010) states that, particularly for teachers who have grown up with testing declarative knowledge, it is difficult to adapt to the

idea of competence orientation and different functions of assessment (summative assessment versus formative assessment). He adds that in mathematics education, conventional procedures, particularly in arithmetics, seem to be objectively assessable. And therefore, he claims, they often turn out to be the only aspect of the subject that is assessed. This is in consistence with Adamina (2010) who describes the situation in science education. He points out that assessment is very often reduced to declarative knowledge. As soon as non-declarative knowledge is assessed, he adds, the criteria of assessment are dominated by formal issues such as length of a talk or layout of slides. He considers this situation unsatisfactory and strongly promotes the idea of assessment as a means of support and guidance: Assessment should help the students to develop self-regulated learning, curiosity, and interest. He adds that the assessment in the classrooms should be better aligned with the educational objectives in the curriculum and with the teaching. Curricular guidelines, assessment and teaching should therefore all be integrated for the lesson planning.

Switzerland has no culture of high-stake large-scale assessment. Instead, summative assessment is at the responsibility of the individual teacher at all school levels with only few exceptions. There are regional large-scale assessments, but they are mostly to evaluate the quality of the school system.

There is little literature on the teachers' assessment practices but the impression emerging from it is that summative assessment is usually done in classical paper-and-pencil tests and often focusses on knowledge. With more innovative forms of assessment such as presentations, the focus typically shifts to formal issues.

3.8.3 Formative assessment

Curricular guidelines

In the curriculum 21 (*Lehrplan 21*, introduced above), formative assessment is given much more weight than in earlier cantonal curricula. Formative assessment is introduced as “accompanying the learning process” (D-EDK, 2014, p. 9) and contrasted to “concluding the learning process” or summative assessment (D-EDK, 2014, p. 10). The instructions on how to enact formative assessment say that formative assessment should support subsequent learning; that it should be provided on an individual basis; and that it should make the current level of achievement transparent but also enclose concrete advice on how to proceed (D-EDK, 2014). With the implementation of the new curriculum, more official documents that focus on assessment are elaborated.

For the *Gymnasium* level, there are no official guidelines on the use of formative assessment.

Formative assessment traditions

According to Vögeli-Mantovani (1999), formative assessment has been an issue in Switzerland since the early eighties (*Reisetagebücher* or *Lerntagebücher*; Ruf & Gallin, 1991) and has been discussed, conceptionally consolidated, and tested particularly at primary school level. The fundamental ideas include student self-assessment and an enhanced awareness of the student's responsibility for their own learning. There are assumptions, but no systematic surveys on formative assessment practices. The above-mentioned author also pointed out the difficulties related to the federal system in Switzerland: The culture of assessment still varies - between different states as well as between different levels of education (Vögeli-Mantovani, 2009). Legal standards as well as educational traditions allow great individuality even at the level of school units (Husfeld, 2009) and at the level of individual classes (Kronig, 2009).

A number of recent school books have incorporated ideas for formative assessment: For example the Math-Buch with assessment materials attuned for different levels of performance, but also MilleFeuille with questions for students to self-check and reflect on their levels of achievement.

Research on formative assessment

There is little research on formative assessment practices in Switzerland. Nevertheless, there is data on the attitude to assessment in the format of oral or written feedback provided by the teachers (instead of the traditional grades which include no guidance for further learning): New ways of assessment such as oral feedback at the end of the semester and learning reports (instead of grades) are highly accepted among teachers. In *Kanton*-wide evaluations (*Kanton* = "state"), over 90% of teachers rated oral feedback by the teachers positive, and more than 75% thought that learning reports on the progress of their students are important (Vögeli-Mantovani, 1999). Similar results were found on the attitude to self-assessment of the students: 90% of teachers acknowledged their value. Comparably high approval was found among parents and students (Vögeli-Mantovani, 1999).

Related to the uptake of more formative assessment, different clusters of challenges have been reported: Traditional views of assessment (such as: Grades as main aim of assessment; mistakes are bad; teacher is fully responsible for assessment) have been mentioned by Rothenbacher (2010). He argues that the problem is not limited to science and mathematics education, but particularly pronounced here since these subjects are thought to be graded more objectively than, for example, essays in language education. Similar to the above-mentioned authors writing about the situation of summative assessment in Switzerland, Jundt (2013) adds that the main purpose of assessment is generally generating grades for school reports, and that the full potentials of assessment for diagnosis and support of student learning are not tapped.

In order to promote formative assessment, Smit (2009) considers the gradual change of culture of assessment and teaching in the schools necessary. This transformation of the teacher's mentality towards assessment as means of enhancing the student's learning should be supported from outside, by the educational system. A second necessity is, according to Smit and Birri (2012), ready-made units including rubrics for assessment which will encourage the teachers to assess complex and therefore often neglected competences.

Smit (2009) shows in his studies that both teachers and students are focussed on grades rather than prospective feedback. Part of the explanation for the result is that not all teachers understand the purpose of formative assessment. Considering formative assessment practices, Smit (2009) regards the lack of knowledge on how to differentiate between several levels of proficiency in the same class as a main flaw on the side of teachers. Another reason that hinders the uptake of more formative assessment is the lack of time. In teacher interviews, teachers reported to have no time to give feedback during classroom hours, and to have no time to develop tools for formative assessment with peer teachers either (Smit, 2009).

The new curriculum for the compulsory school levels puts emphasis on the role and importance of formative assessment, whereas no guidelines exist for the upper secondary school level (*Gymnasium*). Formative assessment strategies are also part of some new school books in various subjects, particularly at primary school level.

Little is known about the formative assessment practices in Switzerland. The picture that emerges is that formative assessment is generally accepted by teachers but the frequency of its use varies a lot between different school levels and individual teachers. This also has to do with the legal standards that allow great individuality at the level of school units and at the level of individual classrooms. Furthermore, assessment practices are generally dominated by summative purposes and some teachers seem to lack abilities and time to employ formative assessment practices.

4 Research questions

Achievement gains associated with formative assessment are among the largest for educational interventions (Black & Wiliam, 1998; Hattie, 2009). From the perspective of the OECD, this “highlights the importance of firmly embedding formative assessment within the broader evaluation and assessment framework and the need to support teachers’ capacity and professionalism in formative assessment” (OECD, 2013, p. 145). In Switzerland, formative assessment activities have been part of teaching practice, but with a high variation between individual teachers and school levels (Vögeli-Mantovani, 1999). With the new curriculum *Lehrplan 21*, an attempt has been taken to stress the importance of formative assessment to enhance student competences (D-EDK, 2014).

But the success of formative assessment policies heavily depends on their effective implementation (Black, 1993; Black & Wiliam, 1998; Stiggins, Griswold & Wikelund, 1989). This is because the quality of formative assessment rests to a high degree on strategies teachers use to elicit evidence of student learning, and on the use of this evidence to shape subsequent instruction and learning (Bell & Cowie, 2001; Heritage, 2010; Herman et al., 2010; Ruiz-Primo et al., 2010).

The overarching aim of this study therefore is to explore how the transfer of formative assessment from the level of national and international educational policy to daily teaching practice could be supported. As laid out above, the successful implementation of an innovation depends on the teachers. This is particularly true in an educational system where teachers have a high autonomy as in Switzerland. The model of professional growth (Clarke & Hollingsworth, 2002) describes changes in teaching practice and was therefore chosen as a theoretical framework for the study. It will be introduced in the following sub-chapter, followed by the research questions.

4.1 A theoretical frame for innovation in teaching: The model of professional growth

The implementation of a relatively uncommonly used approach into regular teaching practice is not an easy endeavour (e.g. Black & Atkin, 1996; Furtak et al., 2008; Smith & Gorard, 2005; Tierney, 2006). It needs teacher change. An empirically grounded framework to study the process of teacher change is the model of teacher professional growth (Clarke & Hollingsworth, 2002). The model describes the “processes by which teachers grow professionally and the conditions that support and promote that growth” (Clarke & Hollingsworth, 2002, p. 947). It suggests that a teacher’s world embraces four domains as pictured in Figure 4: The personal domain; the domain of teaching practice; the domain of consequence; and the external domain. Change occurs “through the mediating processes of ‘reflection’ and ‘enactment’” (Clarke & Hollingsworth, 2002, p. 950) by which one domain affects the other domains.

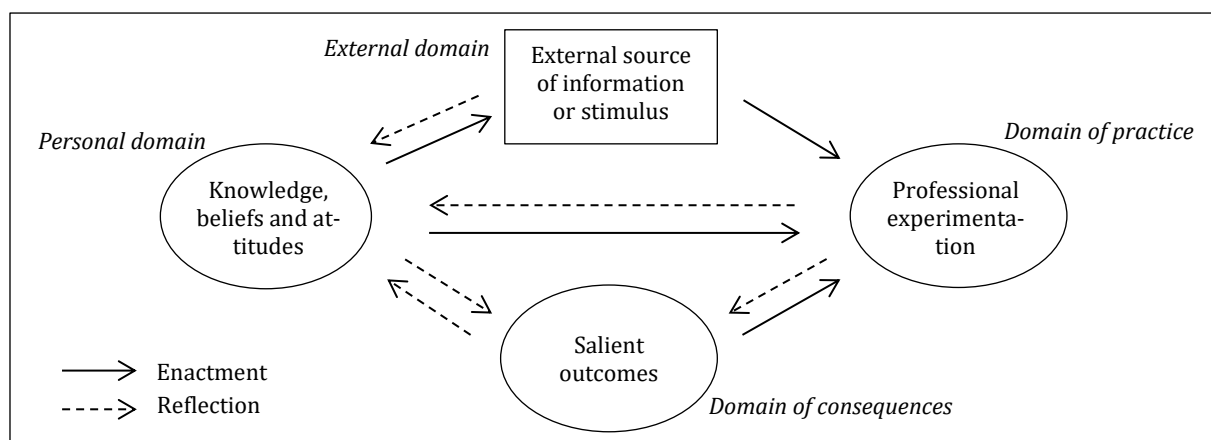


Figure 4: Model of professional growth. From Clarke & Hollingsworth, 2002, p. 951.

Applying the model to the study’s context as summarized in the preceding theory chapter, the external domain is represented, for example, by the promotion of formative assessment in the Swiss curriculum for the

compulsory school levels (see sub-chapter 3.8) but also in international position papers which act as stimuli. The personal domain is represented by teachers' knowledge and beliefs about student learning and assessment (see sub-chapter 3.5). The domain of practice is reflected in teachers' enactments of different formative assessment methods (theoretical background in sub-chapters 3.2 and 3.4). The domain of consequence is, for example, represented by the scientific studies that describe the effect of formative assessment on student achievement (see sub-chapter 3.3).

Linking the subsequent empirical chapters to the model of professional growth, different domains will be explored: In the study, twenty teachers trialled formative assessment methods in their inquiry-based science teaching at primary and at upper secondary school level in Switzerland over the course of three semesters. The personal domain from the model of Clarke & Hollingsworth contains big constructs like knowledge, beliefs and attitudes. It is not easy to capture these in a thorough manner. However, aspects of the personal domain, the teachers' understanding of formative assessment (see research question 1) and their formative assessment self-efficacy (see research question 4.2) will be explored. The domain of practice is probably the part of the model from Clarke & Hollingsworth that was investigated in most detail in the study – the results on research question 2 provide insights into the teachers' professional experimentation. The domain of consequences is well-described in the literature on formative assessment – yet measuring the effect of different formative assessment practices on the students' achievement was never an aim of the study. Looking at the data of the study, the teachers' and the students' evaluations of the different formative assessment methods (research question 3) can be considered outcomes, if 'outcomes' are not restricted to objectively measurable effects on student learning but also includes subjective impressions from teachers and students. The model from Clarke & Hollingsworth does not only consist of domains but also of relations between these domains. It is distinguished between enactment and reflection. Research question 4 on the changes in the teachers' understanding and implementations throughout the study will provide insights into the interplays between the different domains of the model of professional growth.

4.2 Introduction of research questions

For the practical part of the study, twenty teachers from two different school levels explored ways to integrate formal formative assessment methods in their inquiry-based science education in a collaboration that lasted for three semesters. The setting was rather open in the sense that the teachers had considerable freedom in their trials. The research questions of the study will be introduced in the following paragraphs.

In research question 1, the teachers' conceptions of formative assessment will be explored. The aim is to identify potential misconceptions as well as aspects of formative assessment that appear particularly relevant to the teachers collaborating in the study. The question relates to aspects of the personal domain (knowledge, beliefs and attitudes) of the model of professional growth introduced in sub-chapter 4.1.

RQ 1: What is the teachers' understanding of formative assessment?

Research question 2 aims at exploring the use of three formal formative assessment methods in the context of inquiry-based education by the teachers collaborating in the study. The three assessment methods are written teacher assessment, peer- assessment, and self-assessment; see sub-chapter 3.4. From the results, interpretations on the use of formative assessment at the respective school levels will be drawn. Furthermore, problematic aspects in the use of the methods (which teachers may be aware of or not) will be identified. The question relates to the domain of practice (professional experimentation) of the model of professional growth introduced in sub-chapter 4.1.

RQ 2: How do the teachers in the study trial formative assessment methods in their inquiry-based science education?

- 2.1 The inquiry units used for trialling the assessment methods
- 2.2 Ways of putting the formative assessment methods into practice
- 2.3 Problems in the trials

Research question 3 will explore the teachers' evaluation of the formative assessment methods trialled in the study. The aim is to identify potential benefits of the assessment methods from the teacher perspective as well as challenges on different levels (in the classroom, on a systemic level, at the level of teaching resources, etc.). In the second part of the question, the perspective of the students will be taken into account. This appears relevant because the students' acceptance of formative assessment is likely to heavily influence the success of respective activities. The question relates to the domain of consequences (salient outcomes) of the model of professional growth introduced in sub-chapter 4.1.

RQ 3: How do the teachers and the students evaluate the formative assessment methods trialled?

- 3.1 Usability of the methods for different school levels as perceived by the teachers
- 3.2 Benefits and challenges of specific assessment methods as mentioned by the teachers
- 3.3 Means of support as mentioned by the teachers
- 3.4 Usability as perceived by the students
- 3.5 Benefits and challenges as mentioned by the students
- 3.6 Means of support as mentioned by the students

In research question 4, the aim is to explore potential changes in the formative assessment practices and perceptions throughout the collaboration in this study where the teachers develop their own assessment. With the small sample sizes, the results on research question 4 are clearly tenuous. Due to the little literature available, it nevertheless appeared legitimate to conduct the respective analyses. The interpretation of the results will be done with caution. Part of this cautious interpretation is that the data will, in some sections, not be analysed separately for the two school levels as for the other research questions. Instead, the teachers will here be considered as one group. The question relates to the model of professional growth introduced in sub-chapter 4.1 by exploring the changes that occur through enactment and reflection.

RQ 4: How does the teachers' understanding and implementation of formative assessment change throughout the collaboration in the study?

4.1 Changes in the understanding of formative assessment

4.2 Changes in the self-efficacy

4.3 Changes in the implementations

4.4 Changes in the importance, benefits and challenges perceived

4.5 Support mechanisms from the collaboration in the study

4.6 Variability of implementations within teachers

4.3 Use of results for generation of hypotheses

This is an explorative study with a small number of participating teachers and the overwhelming portion of data analysis will be qualitative. The results are therefore used for the generation of hypotheses.

From the results, two sets of hypotheses are deduced: The first set of hypotheses focusses on the conditions and measures of support for the implementation of formative assessment practices in Switzerland. This set of hypotheses (H1 – H6) is based on the results to research questions 1, 2, and 3. The second set of hypotheses focusses on the implementation behaviours of teachers. That set of hypotheses (H7 – H8) is based on the results to research question 4. The exact connections between the research questions and the hypotheses derived from the respective results can be found in Table 4.

Table 4: Connections between research questions from sub-chapter 4.2 and the hypotheses derived from the results in sub-chapters 8.5 and 8.6.

RQ 1: The teachers' understanding of formative assessment		Conditions and measures of support for the implementation of formative assessment practices in Switzerland	H1: Teacher concepts and misconceptions of formative assessment
RQ 2: Trials	2.1 Inquiry units 2.2 Formative assessment 2.3 Problems		H2: Teacher attitudes towards formative assessment H3: Aims pursued with formative assessment
RQ 3: Evaluation of methods	3.1 Usability 3.2 Benefits and challenges 3.3 Means of support 3.4-6 Student perspective		H4: Formative assessment practices for the school levels explored H5: Problem areas for the uptake of formative assessment H6: Support for the uptake of formative assessment practices
RQ 4: Changes throughout the collaboration in the study	4.1 – 4.4 Changes 4.5 Support mechanisms in the study 4.6 Variability within teachers	Implementation behaviour of teachers	H7: Effects of the study on the teachers' understanding and practices H8: Implementer types

5 Methods

In this chapter, the methods of the study will be laid out. Sub-chapter 5.1 will introduce the setting of the study and the links to the ASSIST-ME project. Sub-chapter 5.2 will provide the details on the participants of the study. In sub-chapter 5.3, the instruments for the data collection and their use will be described. The subsequent sub-chapters 5.4 and 5.5 will then introduce the details on the data analysis.

5.1 Setting

As explained in the project overview in chapter 2, this study was integrated in the ASSIST-ME project. There was a large overlap in the design: The participants, the time of data collection and some of the methods of data collection were the same. The basic principle for both the ASSIST-ME project (Dolin, 2012) and for this study was that 20 teachers from primary school and from upper secondary school trialled formative assessment methods during three semesters of collaboration. This collaboration with the teachers started in August 2014 and lasted until January 2016, so it involved the fall term 2014; the spring term 2015; and the fall term 2015.

At the beginning of the first semester, two introductory meetings of two and a half hours each were set up for the teachers of both school levels together. During these meetings, the conceptual understanding of formative assessment in the study was presented and discussed amongst the teachers. The formative assessment methods used in the study (written teacher assessment; peer-assessment; self-assessment; see sub-chapter 3.4) were introduced with concrete examples of inquiry-based science units. Questions from the teachers were answered. Both the theoretical explanations and the examples were given to the teachers in the so-called 'manual', a booklet of 35 pages. Furthermore, the teachers had access to articles from teacher journals and to videos on formative assessment and on inquiry teaching provided through a shared dropbox folder.

The teachers were then asked to choose at least one of the formal formative assessment methods (written teacher assessment; peer-assessment; self-assessment; see sub-chapter 3.4), to integrate it in any of their inquiry units in their own classrooms and use the assessment method to assess inquiry-relevant competences (see sub-chapter 3.1). The teachers were free to trial the assessment methods at any time during the first semester of collaboration. They were introduced to the methods of data collection and particularly to their duties relating to the documentation of their trials (see sub-chapter 5.3). The teachers were encouraged to contact the Center for Science and Technology Education at PH FHNW if they had the impression that they needed advice or support.

At the end of the first semester, another meeting of two and a half hours for all teachers was organised in order to discuss the experiences with the first round of trials. It was a meeting for the teachers from both school levels, but the discussions were held in smaller groups according to the school level taught at. In this end-of-semester meeting, the teachers brought up more questions which were discussed. The teachers were encouraged to exchange their teaching materials from their units with formative assessment: Via the shared dropbox folder; through direct communication during the meeting; or by any other means.

In the second and in the third semester of implementation, introductory meetings of two and a half hours at the beginning of both semesters were held. In these, more examples of the use of the three formative assessment methods were discussed. Furthermore, there were short input presentations and longer group works within and across school levels on aspects of inquiry and on the curriculum in Switzerland. These two introductory meetings were also used to set additional guidelines regarding the trials: These additional guidelines were, firstly, that in the trials, the criteria of assessment had to be implicitly or explicitly clear for the students. Secondly, the students must have the opportunity to make use of the feedback received in the formative assessment activity: Either in the same unit or at another occasion.

The teachers were asked to trial, similarly to the first semester, at least one of the formal formative assessment methods in any of their inquiry units in their own classrooms. They were also reminded of the necessary documentation of these trials and they were asked to contact the Center for Science and Technology Education at PH FHNW in case of any doubt or if support was needed. The two end-of-semester meetings to exchange experiences were similar to the respective meeting in the first semester.

Relating this study to the different strategies to implementing formative assessment methods introduced in sub-chapter 3.7, it resembles the bottom-up approaches from 3.7.2 where teachers developed, with the help of researchers or autonomously, their own assessment strategies rather than receiving ready-made materials for usage.

5.2 Participants

For both the ASSIST-ME project and for this study, 20 engaged science teachers were searched. Apart from the high commitment for teaching, emphasis was put on an even distribution in gender and years of teaching experience. The teachers had to represent at least two different school levels and different subjects for the needs of the ASSIST-ME project. In Switzerland, the primary and the upper secondary school levels were chosen because of their different socialisations from more pedagogically-oriented backgrounds (primary school teachers) to more subject-oriented backgrounds (upper secondary school teachers). The primary school teachers taught integrated science; the upper secondary school teachers taught physics, chemistry, or biology.

The teachers were contacted and asked for collaboration individually: Some of them had collaborated in projects of the Center for Science and Technology Education before, others were found via lecturers from the School of Teacher Education PH FHNW. With the first contact, the teachers were told that they would be choosing and trialling assessment methods in their normal teaching with their classes; that there would be about ten meetings to attend over the course of three semesters; and that they would be paid for documenting their trials, for filling out questionnaires, and for attending the meetings.

Table 5 below displays the characteristics of the teachers from primary school who participated in the study. The two teachers who are marked with an asterisk (*) left the project after two semesters for personal reasons. Apart from the information displayed in the table, the teachers were also asked about their own educational background. They all had a teaching degree for their school level which involves courses in general educational science as well as in science education.

Table 5: Participants of the study teaching at primary school (N=9). Teachers marked with an asterisk left the project after two semesters.

ID	Teaching experience [years]	Gender	Subjects taught	School level [grades]
P1*	0-4	f	integrated science, mathematics, handicrafts, sports	1-3
P2	0-4	m	integrated science, mathematics, german, french	4-6
P3	0-4	f	integrated science, mathematics	4-6
P4	5-10	f	integrated science, mathematics, german, french	4-6
P5*	5-10	f	integrated science, mathematics	1-6
P6	11-20	f	integrated science	1-6
P7	11-20	m	integrated science, mathematics, music, sports	4-6
P8	>20	f	integrated science, mathematics	4-6
P9	>20	m	integrated science, mathematics, music, sports	4-6

Table 6 below displays the characteristics of the teachers who participated in the ASSIST-ME project from upper secondary school. The teachers who are marked with two asterisks (**) collaborated with Olia Tsivitanidou for one (S8) or two (S9; S10) semesters. The trials that emerged from that collaboration are not included in this study. Instead, parts of them are documented in Tsivitanidou and Labudde (2016). Apart from the information displayed in the table, the teachers were also asked about their own educational background. They all had a teaching degree for their school level and subject(s) which involves courses in general educational science as well as in science education.

Table 6: Participants of the study teaching at upper secondary school (N=11). Teachers marked with two asterisks collaborated with Olia Tsivitanidou for one (S8) or two (S9; S10) semesters. The trials that emerged from that collaboration are not included in this study.

ID	Teaching experience [years]	Gender	Subjects taught	School level [grades]
S1	0-4	f	physics, mathematics	7-12
S2	0-4	m	chemistry, biology	7-12
S3	0-4	f	chemistry	9-12
S4	5-10	m	physics, mathematics	9-12
S5	5-10	m	physics, mathematics	9-12
S6	5-10	m	chemistry, mathematics	7-12
S7	11-20	f	biology	7-12
S8**	11-20	f	biology	9-12
S9**	11-20	f	biology	7-12
S10**	>20	m	physics	9-12
S11	>20	m	biology	9-12

5.3 Data collection

The data collection took place from August 2014 until January 2016 in three rounds of one semester each. Some data was only collected at the beginning and at the end of the collaboration with the teachers (e.g. the teacher profile questionnaire), whereas other data was collected in every round (e.g. group discussions with the teachers). Therefore, the data collection is visualized in two figures: Figure 5 provides the broader overview of the whole data collection and displays the data that was collected at the beginning and at the end of the collaboration with the teachers only. Figure 6 provides a closer insight into the data that was collected in every round of implementation. Details on the different types of data will be provided in the subsequent sections 5.3.1 – 5.3.10.

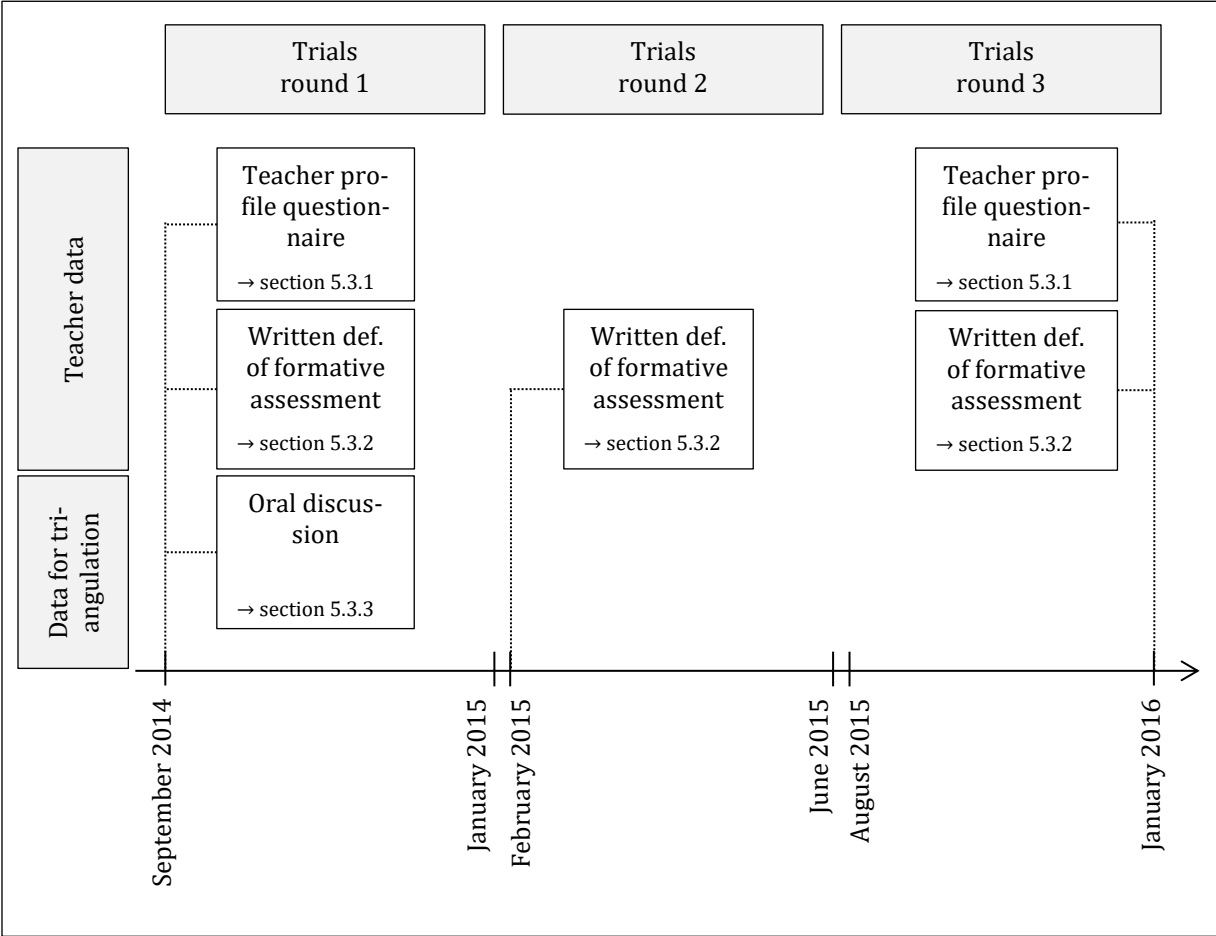


Figure 5: Overview of data collection.

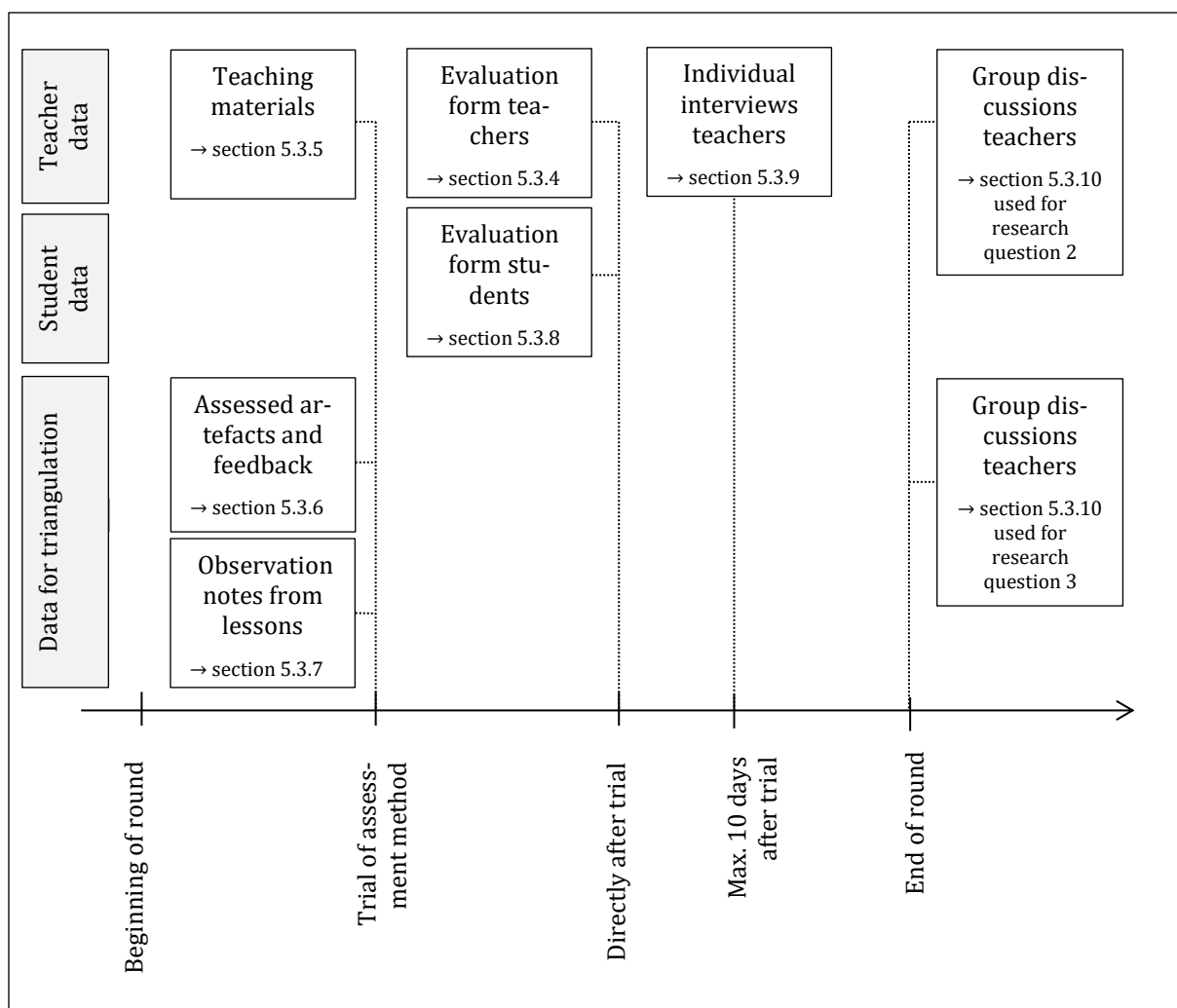


Figure 6: Data collected in every round of implementation.

The data collection was planned considering four guiding aspects: The research questions, methodical reasons, the ASSIST-ME project guidelines, and practical reasons. These aspects will be introduced in more detail below.

For the planning of the data collection, the research questions had to be taken into account first and foremost: In order to investigate the four research questions from chapter 4, the focus of data collection had to be on the teacher perspective. In more detail, the teachers' understanding of formative assessment had to be captured (for research question 1); the reconstruction of the teachers' trials had to be possible (for research question 2); the teachers' perceptions of advantages and challenges had to be captured (for research question 3); and possible changes throughout the collaboration in the study had to be reconstructed (for research question 4).

Secondly, the data collection was influenced by methodical reasons: Since the majority of the research questions were investigated with qualitative methodology, triangulation appeared particularly relevant.

Thirdly, the ASSIST-ME project guidelines were taken into account when planning the data collection: The study was situated within the ASSIST-ME project, as laid out in chapter 2 and sub-chapter 5.1. In order to use synergies, some of the instruments used for data collection in the ASSIST-ME project were supplemented with additional items so that they could also feed into the research questions of this study. An example of such an instrument is the teacher profile questionnaire (see section 5.3.1).

Lastly, practical reasons influenced the data collection: There are many research projects and professional development programmes for teachers in Switzerland. Furthermore, the teachers' workload is generally high. So to make sure that participants for the study could be found, the teachers had to be ensured that their effort for data collection in the study would be manageable.

The reasoning for the concrete instruments of data collection can be found in the following sections.

5.3.1 Teacher profile questionnaire

The questionnaire was developed in English by the Danish partners of the ASSIST-ME project and afterwards translated to German. It contained, on the one hand, questions on the demography of the teachers: The teaching experience, the subjects and the school levels taught at. On the other hand, the questionnaire also included items on the teachers' formative and summative assessment practices, their impression of the importance of formative and summative assessment as well as their self-efficacy in formative assessment. The items that were used for this study (41 items) can be found in appendix A1.

The questionnaire was not only filled out by the teachers participating in the study but also by a control group. Each teacher was asked to find a peer science teacher from the same school and the same school level to form this control group.

The purpose of this instrument within the study was, firstly, to describe the participants (see sub-chapter 5.2), and secondly to contribute to research question 4 which investigated possible changes of the teachers' beliefs and practices throughout the collaboration in the study. The instrument was chosen because of the synergies with the ASSIST-ME project.

Table 7 provides an overview of the data collected in the teacher profile questionnaire. The data was collected in an electronic survey administered by the Danish partners of the ASSIST-ME project with SurveyXact and transferred into an excel file. The results were collected at the beginning of the project (September 2014) and at the end of the project (January 2016). For the analysis in the context of research question 4, only the teachers who filled out the survey at both measurement points were included.

Table 7: Data from the teacher profile questionnaire

	September 2014	January 2016
Participants of the study	20 filled-out questionnaires in total - 9 questionnaires from primary school teachers - 11 questionnaires from upper secondary teachers	16 filled-out questionnaires in total - 6 questionnaires from primary school teachers - 10 questionnaires from upper secondary teachers
Control group	18 filled-out questionnaires in total - 5 questionnaires from primary school teachers - 13 questionnaires from upper secondary teachers	13 filled-out questionnaires in total - 5 questionnaires from primary school teachers - 8 questionnaires from upper secondary teachers

5.3.2 Written definition of formative assessment from teachers

Three times throughout the course of the study (at the beginning of the first meeting in September 2014; in the meeting in May 2015; at the end of the last meeting in January 2016), all teachers were asked to explain what formative assessment means in their understanding. The teachers were asked to do that in a written form during the meetings in no more than 10 minutes; the template can be found in attachment A2. This task was completed anonymously but the three explanations of every teacher could be linked by a code (initials of the mother and year of birth of the mother). Apart from this code, the teachers were also asked to indicate the school level they taught at. Since not all teachers were present in all meetings in which the

data collection took place, the data set is not complete (see Table 8). The writings were transcribed into a word file for further analysis.

The purpose of this instrument within the study was to contribute to research question 1 and question 4.1 which investigated the teachers' understanding of 'formative assessment'. Only teachers who wrote definitions at two or three measurement points were included in the respective analyses. The instrument was chosen for practical reasons: It was a quick and easy method to capture data during the meetings with all teachers (see sub-chapter 5.1). The analysis of relatively short, written texts was expected to be manageable.

Table 8: Data from the written definition task.

September 2014	May 2015	January 2016
20 written definitions in total - 9 definitions from primary school teachers - 11 definitions from upper secondary school teachers	15 written definitions in total - 7 definitions from primary school teachers - 8 definitions from upper secondary school teachers	16 written definitions in total - 6 definitions from primary school teachers - 10 definitions from upper secondary school teachers

5.3.3 Oral discussion on the meaning of the term 'formative assessment'

At the very beginning of the collaboration with the teachers, in September 2014, the meaning of the term 'formative assessment' was discussed during 20 minutes in groups of 5-8 teachers with the groups formed according to school level and subject. This discussion took place right after the teachers had been asked to define the term in a written form. The discussions were audiotaped and transcribed.

The purpose of this instrument within the study was to contribute to research question 1 (teachers' understanding of what 'formative assessment' is). The instrument was chosen for both methodical and for practical reasons: On the one hand, it allowed for triangulation in research question 1. On the other hand, the effort for data collection was small as the meeting with all teachers took place anyways.

5.3.4 Evaluation form for teachers

In every semester of implementation, all teachers were asked to fill out a questionnaire with open questions on their inquiry-based unit with formative assessment (what they trialled) and on the evaluation of it of (the benefits and the challenges they perceived). This evaluation form also included the teaching plans. The template of the evaluation form can be found in attachment A3. The teachers were handed out the evaluation forms both as a hardcopy and in a digital version so that everyone could answer the question in the way preferred. The teachers were asked to fill out the evaluation form right after the implementation. An overview of the resulting data can be found in Table 9 below. The answers of the teachers were transcribed into an electronical table for further analysis.

The purpose of this instrument within the study was to contribute to research questions 2 (reconstruction of the trials) and 3 (teacher's perceptions of benefits and challenges). The instrument was chosen because the teachers had to fill out a similar form for the ASSIST-ME project, so by adding questions for the purpose of this study, respective synergies could be used. As the formative assessment methods trialled did not include on-the-fly or similarly spontaneous methods but mostly written interactions, videotaping of lessons did not appear appropriate.

Table 9: Data from the evaluation forms for teachers.

1 st semester of implementation	2 nd semester of implementation	3 rd semester of implementation
15 evaluation forms in total - 9 evaluation forms from primary school teachers - 6 evaluation forms from upper secondary school teachers	16 evaluation forms in total - 7 evaluation forms from primary school teachers - 9 evaluation forms from upper secondary school teachers	12 evaluation forms in total - 2 evaluation forms from primary school teachers - 10 evaluation forms from upper secondary school teachers

5.3.5 Teaching materials

In every semester of implementation, all teachers were asked to hand in the teaching materials of the inquiry-based unit with formative assessment. These included photocopies and work sheets for students, assessment rubrics, and similar materials. The teachers handed in these materials as hardcopies or in a digital version. An overview of the resulting data can be found in Table 10 below.

The purpose of this instrument within the study was to contribute to research question 2 (reconstruction of the trials). The instrument was chosen for both methodical and for practical reasons: On the one hand, it allowed for a more detailed reconstruction of the teachers' trials. On the other hand, providing the teaching materials was no additional effort for the teachers participating in the study.

Table 10: Teaching materials.

1 st semester of implementation	2 nd semester of implementation	3 rd semester of implementation
15 sets of teaching materials in total - 9 sets of teaching materials from primary school teachers - 6 sets of teaching materials from upper secondary school teachers	16 sets of teaching materials in total - 7 sets of teaching materials from primary school teachers - 9 sets of teaching materials from upper secondary school teachers	12 sets of teaching materials in total - 2 sets of teaching materials from primary school teachers - 10 sets of teaching materials from upper secondary school teachers

5.3.6 Assessed student artefacts and feedback provided

In the second and in the third semester of the project, student artefacts such as lab reports and corresponding feedback were collected from a small number of classes. An overview of the collected data can be found in Table 11. The original data used in the classes were photocopied and afterwards transcribed for further analysis.

The purpose of this instrument within the study was to contribute to research question 2 (reconstruction of the trials). The instrument was chosen for both methodical reasons and for reasons related to the ASSIST-ME project: On the one hand, it allowed for triangulation. On the other hand, a number of teachers were requested to provide student artefacts and feedback for analysis within the ASSIST-ME project anyways. So synergies could be used.

Table 11: Assessed student artefacts and corresponding feedback.

1 st semester of implementation	2 nd semester of implementation	3 rd semester of implementation
---	121 sets of student artefacts and feedbacks in total <ul style="list-style-type: none"> - Student artefacts and teacher feedback from a primary class (N=21 students) in integrated science - Student artefacts from a primary class (N=24 students) in integrated science - Student artefacts and teacher feedback from an upper secondary class (N=17 students) in biology - Student artefacts and teacher feedback from an upper secondary class (N=23 students) in physics - Student artefacts and peer-assessment from an upper secondary class (N=19 students) in physics - Student artefacts and self-assessment from an upper secondary class (N=17 students) in biology 	62 sets of student artefacts and feedbacks in total <ul style="list-style-type: none"> - Student artefacts and peer-assessment from an upper secondary class (N=23 students) in physics - Student artefacts and peer-assessment from an upper secondary class (N=19 students) in physics - Student artefacts and peer-assessment from an upper secondary class (N=20 students) in physics

5.3.7 Observation notes from lessons visited

In the second and in the third semester of the project, a number of teachers who agreed on it were visited in their classes when using one of the formative assessment methods in inquiry-based education. The observations were documented with general notes of what was going on in the classes as well as more specific notes on the teachers' instructions regarding the formative assessment activities and both the teachers' and the students' (re)actions on the formative assessment. An overview of the collected data can be found in Table 12. The observational notes were transcribed into a word file.

The purpose of this instrument within the study was to contribute to research question 2 (reconstruction of the trials). The instrument was chosen for both methodical and for practical reasons: On the one hand, it allowed for a more detailed reconstruction of the teachers' trials. On the other hand, having a person visiting lessons was no additional effort for the teachers participating in the study.

Table 12: Observation notes from lessons visited.

1 st semester of implementation	2 nd semester of implementation	3 rd semester of implementation
---	Observational notes from 8 lessons in total <ul style="list-style-type: none"> - Observational notes from a unit with 4 physics lessons where written comments provided by the teacher were used - Observational notes from a unit with 2 integrated science lessons were written comments provided by the teacher were used. - Observational notes from a unit with 2 physics lessons where peer-assessment was used 	Observational notes from 4 lessons in total <ul style="list-style-type: none"> - Observational notes from a unit with 4 physics lessons where peer-assessment was used

5.3.8 Evaluation form for students

In the second and in the third semester, a small number of classes of teachers who participated in the study were asked to fill out an evaluation form after the trial of a formative assessment activity. The form included one closed and six open questions asking about the benefits, usability, difficulties and challenges with the assessment method trialled. The template of the evaluation form can be found in attachment A4. The students remained anonymous.

The evaluation form for students was handed out by the teachers as a hard copy and filled out during classroom hours right after the trials. An overview of the resulting data can be found in Table 13 below. The answers of the students were transcribed into an electronic table for further analysis.

The purpose of this instrument within the study was to contribute to research questions 3.4 – 3.6 (students' perspective on formative assessment). The instrument was chosen for reasons concerning both the research questions and methodical issues: The perspective of the students was considered relevant for a successful implementation of an innovative approach in the classroom, it was therefore crucial for the study to not only collect data on the teachers' perspective. This appeared particularly true in the case of methods with a high student involvement. The data on the student perspective was therefore restricted to peer-assessment. The method of data collection was chosen for practical reasons: The sample was restricted to upper secondary school students only because written data collection is fast with them.

Table 13: Data from the evaluation forms for students.

1 st semester of implementation	2 nd semester of implementation	3 rd semester of implementation
Evaluation form from three upper secondary classes (N=63) in physics working on the same implementation (by the same teacher) with peer-assessment	Evaluation form from an upper secondary class (N=19) in physics working with both peer-assessment and written teacher assessment	Evaluation form from an upper secondary class (N=21) in physics working with peer-assessment

5.3.9 Individual interviews with teachers

With a small number of teachers, semi-structured interviews of about 40 minutes were conducted. The interviews took place within 10 days after the end of the trials. The individual interviews included questions on the preparation, the conduction and the evaluation of the implementation as well as on the teachers' formative assessment practices in general. There were also questions on the benefits of the collaboration in the study. The interview guide can be found in attachment A5.

The teachers were selected so that they covered the different subjects taught, the different school levels, gender and teaching experience. An overview of the resulting data can be found in Table 14 below. The interviews were conducted in dialect and audiotaped. They were transcribed in standard German for further analysis.

The purpose of this instrument within the study was to contribute to research questions 2, 3, and 4.5. The instrument was chosen for methodical reasons: It allowed for a more detailed picture of the teachers' perspectives. Since individual interviews are time-consuming, they were not conducted with all teachers but with a limited selection of the participants of the study.

Table 14: Data from the individual interviews with teachers.

1 st semester of implementation	2 nd semester of implementation	3 rd semester of implementation
7 interviews in total - 3 individual interviews with primary school teachers - 4 interviews with upper secondary school teachers	6 interviews in total - 3 interviews with primary school teachers - 3 interviews with upper secondary school teachers	3 interviews in total - 3 interviews with upper secondary school teachers

5.3.10 Group discussions with teachers

At the end of every semester, a meeting with all teachers participating in the study took place. In that meeting, group discussions of about 45 minutes duration took place on the results and experience from the respective round of implementation. So the teachers, on the one hand, spoke about the trials but on the other hand also about the challenges and the advantages that they perceived. The exact questions discussed can be found in attachment A6. The groups consisted of 5-8 teachers, the groups were formed according to the different school levels and subjects taught. An overview of the resulting data can be found in Table 15 below. The group discussions were conducted in dialect and audiotaped. For further analysis, the group discussions were transcribed and translated to standard German.

The purpose of this instrument within the study was to contribute to research question 2 (reconstruction of the trials) and research question 3 (teachers' perspective on benefits and challenges of formative assessment methods). The instrument was chosen for both methodical and for practical reasons: On the one hand, it provided additional data for research question 2 and it allowed for triangulation in research question 3. On the other hand, the effort for data collection was small as the meeting with all teachers took place anyways.

Table 15: Data from the group discussions with the teachers.

1 st semester of implementation (January 7 th 2015)	2 nd semester of implementation (May 27 th 2015)	3 rd semester of implementation (January 5 th 2016)
<p>3 discussions with 18 teachers in total</p> <ul style="list-style-type: none"> - Discussion 1: 6 integrated science teachers from primary school - Discussion 2: 5 biology teachers from upper secondary school - Discussion 3: 7 physics and chemistry teachers from upper secondary school 	<p>3 discussions with 17 teachers in total</p> <ul style="list-style-type: none"> - Discussion 4: 7 integrated science teachers from primary school - Discussion 5: 5 biology teachers from upper secondary school - Discussion 6: 5 physics and chemistry teachers from upper secondary school 	<p>3 discussions with 16 teachers in total</p> <ul style="list-style-type: none"> - Discussion 7: 5 integrated science teachers from primary school - Discussion 8: 5 biology teachers from upper secondary school - Discussion 9: 6 physics and chemistry teachers from upper secondary school

5.4 Selection of cases for analysis

In the study, 20 teachers from primary and from upper secondary school level implemented different formative assessment methods in their inquiry teaching. Over the course of three semesters, this resulted in 53 cases for this study (some dropouts and some cases analysed by Olia Tsivitanidou; therefore not $3 \times 20 = 60$ trials).

These cases were triaged according to four criteria (C1 – C4):

- C1) Did any trial take place (e.g. did the teacher try out any formative assessment method)?
Criterion spelled out in section 5.4.1.
- C2) Was the trial sufficiently documented to be analysed according to the subsequent aspects 3) and 4)?
Criterion spelled out in section 5.4.2.
- C3) Did the trial take place in the context of an inquiry unit?
Criterion spelled out in section 5.4.3.
- C4) Did the trial involve a formative assessment method?
Criterion spelled out in section 5.4.4.

5.4.1 C1: Conduction of trials

In order to fulfil the criterion on the conduction of a trial, there must be at least one documented sign of the respective trial (e.g. the trial being told about in the group discussions).

5.4.2 C2: Documentation of trials

The trials were documented with evaluation forms, lesson plans, teaching materials, records of group discussions and individual interviews. In order to fulfil the criterion related to the documentation of trials, it must be possible to evaluate the trials in terms of the two subsequent criteria in sections 5.4.3. and 5.4.4 based on the documentations.

5.4.3 C3: Indicators related to inquiry units

A number of indicators were defined in order to decide whether the trialed unit was inquiry-based. These indicators are:

- C3.1) The trialed unit is open in at least one of the dimensions as defined in Priemer (2011): Content; strategy of investigation; methods applied; number of solutions; number of ways to come to a solution; phase in experimentation (see section 3.1.3).
- C3.2) The trialed unit includes at least one inquiry activity as defined in Bell et al. (2010, see section 3.1.1).
- C3.3) The competences that are formatively assessed are domain-specific or transversal competences that are ascribed to inquiry-based science education as defined in section 3.1.2.

5.4.4 C4: Indicators related to formal formative assessment

A number of indicators were defined in order to decide whether the trialed unit contains formative assessment. These indicators are:

- C4.1) The expectations must be clear to both the teacher and the students from the beginning of the unit (see section 3.2.2). This can be explicitly, for example with learning goals, or implicitly, for example in lab sessions where the assessment criteria remain the same throughout the semester and are therefore not repeated constantly.
- C4.2) Timing and procedure of diagnosis (see section 3.2.2) must be planned and clear to both the students and the teacher.
- C4.3) Students must receive the results of that diagnosis in the form of an individual feedback (see section 3.2.2) on their own piece of work.
- C4.4) Students must have the opportunity to use the feedback (see section 3.2.2) either by revising their draft version of artefact or by applying the feedback in a new, similar situation.

C4.5) Since the focus of this study is on a defined selection of formal formative assessment methods; an additional criterion was defined: The formative assessment activities must be assignable to one of the formal formative assessment methods described in the theory part in sub-chapter 3.4.

5.4.5 Cases selected for analysis

Attachment A7 provides an overview of all cases. The cases have been classified depending on whether they fulfil the criteria formulated in sections 5.4.1 to 5.4.4. The resulting number of cases for each group is displayed in Table 16. Depending on the research question, different groups of cases were included in the analysis.

Table 16: Overview of cases.

Characteristics of cases	Frequency
Trial of a formal formative assessment method (according to criteria C4.1 – C4.5) in the context of inquiry (according to the criteria C3.1 – C3.3)	34 cases (14 primary and 20 upper secondary cases)
Trial with no inquiry or no formal formative assessment	9 cases (4 primary and 5 upper secondary cases)
No trial or trial not sufficiently documented	10 cases (7 primary and 3 upper secondary cases)
Total	53 cases

5.5 Data analysis

Most of the data collected in this study was analysed using qualitative content analysis (Mayring, 1994; 2010): The written definitions of formative assessment from the teachers; the teaching materials; parts of the evaluation forms from the teachers; the individual interviews with the teachers; the group discussions with the teachers and the evaluation forms from the students. However, for some sub-questions of research question 4, quantitative analyses were performed. A more detailed insight to the procedure of data analysis and the selection of materials will be provided in the following sub-chapters. The sub-chapters will be structured along the research questions.

5.5.1 Data analysis for research question 1

In research question 1, the teachers' understanding of formative assessment was investigated. For this, the teachers' written definitions of 'formative assessment' (see section 5.3.2) were analysed. Both inductive and deductive coding was used to develop the respective coding system. The data were then analysed using qualitative content analysis (Mayring, 1994; 2010). The quality of data analysis was ensured by double-coding a portion of the data and by triangulation with other data from the study.

Procedure of data analysis for research question 1

The coding system was inductively developed from the teachers' written definitions but the references in the literature were taken into account as well, so both the inductive and the deductive approach were applied. The units of coding were single words. All codes were developed in German and afterwards translated. The coding system was evaluated by double-coding 20% of the data (11 definitions out of 51). The interrater-reliability, measured with Cohen's Kappa, was $\kappa=0.89$. Landis and Koch (1977) consider values ≥ 0.81 as almost perfect agreement. For triangulation, an oral discussion at the first meeting with all teachers (see section 5.3.3) focussed on what formative assessment is. That discussion had taken place immediately after the first written definition task. The triangulation showed that the coding system developed from the written definitions was also applicable to the transcript of the oral discussion.

Coding frame for research question 1

Research question 1 was answered using qualitative data analysis. Table 17 provides an overview of the content of this research question, the dimensions of the coding frame and the data analysed for research question 1. The coding instructions with descriptions and examples for each category can be found in appendix A8.

Table 17: Coding frame for the first research question.

Research question(s)	Dimensions and categories of the coding frame (highest hierarchical levels only)	Coded data
1. Teachers' understandings of formative assessment	<ul style="list-style-type: none"> - Elements ascribed to formative assessment in the literature (deductive development of codes) - Elements ascribed to assessment in the literature (deductive development of codes) - Other elements (inductive development of codes) - Examples of assessment methods (inductive development of codes) 	<ul style="list-style-type: none"> - Written definitions of FA from teachers (see section 5.3.2) Triangulation: - Oral discussion (see section 5.3.3)

5.5.2 Data analysis for research question 2

In research question 2, the teachers' trials were described and analysed. For this, deductive coding was applied to the teaching plans within the evaluation forms, the teaching materials, the transcripts of the individual interviews and the group discussions. The results will be displayed using descriptive statistics. The quality of data analysis was ensured by double-coding a portion of the data and by triangulation with other data from the study. Details will be provided below.

Procedure of data analysis for research question 2

For research question 2, both the cases that match the criteria as specified in sub-chapter 5.4 but also those that do not match these criteria were analysed: The cases that match the criteria were used to answer research questions 2.1 and 2.2; the cases that do not match these criteria were used to answer question 2.3.

The codes for the coding system in research question 2 were derived from the literature as laid out in chapter 3. The data coded consisted of the teaching plans which were included in the teacher evaluation form (see section 5.3.4), the teaching materials (see section 5.3.5), and the transcripts of the individual interviews (see section 5.3.9) and group discussions (see section 5.3.10). The reliability of the coding was evaluated by double-coding 18% of the data (10 cases out of 54 cases). The interrater-reliability, measured using Cohen's Kappa, was $\kappa=0.83$. Landis and Koch (1977) consider values ≥ 0.81 as almost perfect agreement. The results were triangulated with the assessed artefacts and feedback (see section 5.3.6) and the observation notes (see section 5.3.7).

Coding frame for research question 2

Table 18 provides an overview of the content of research question 2, the dimensions of the coding frame and the data analysed. The coding instructions with descriptions and examples for each category can be found in appendix A9.

Table 18: Coding frame for research question 2.

Research question	Dimensions and categories of the coding frame (highest hierarchical levels only)	Coded data
2.1 Description of the inquiry units used in the trials	<ul style="list-style-type: none"> - Dimension(s) of openness in the inquiry units (deductive coding following Priemer, 2011, see section 3.1.3) - Inquiry activities in the units (deductive coding following Bell et al., 2010, and OECD, 2005b; see section 3.1.1) - Competences assessed (deductive coding following Bell et al., 2010, and OECD, 2005b; see section 3.1.2) 	<ul style="list-style-type: none"> - Teaching plans (in the teacher evaluation form; see section 5.3.4) - Teaching materials (see section 5.3.5) - Individual interviews with teachers (see section 5.3.9) - Group discussions (see section 5.3.10)
2.2 Description of the formative assessment activities trialled	<ul style="list-style-type: none"> - Communication of criteria (deductive coding, see section 3.2.2) - Data sources for diagnosis (deductive coding, see section 3.2.2) - Assessment methods (deductive coding, see sub-chapter 3.4) - Means of engaging with the feedback (deductive coding, see section 3.2.2) - Cycle length (deductive development of codes following Wiliam, 2010, see section 3.2.4) 	Triangulation: <ul style="list-style-type: none"> - Assessed artefacts and feedback (see section 5.3.6) - Observation notes (see section 5.3.7)

Table 18 cont.: Coding frame for research question 2.

Research question	Dimensions and categories of the coding frame (highest hierarchical levels only)	Coded data
2.3 Challenges in the trials	<ul style="list-style-type: none"> - Conduction of trials (see sub-chapter 5.4) - Sufficient documentation of trials (see sub-chapter 5.4) - Inquiry-based nature of trials (deductive development of codes, see sub-chapter 5.4) - Formative assessment in trial (deductive development of codes, see sub-chapter 5.4) 	<ul style="list-style-type: none"> - Teaching plans (in the teacher evaluation form; see section 5.3.4) - Teaching materials (see section 5.3.5) - Individual interviews with teachers (see section 5.3.9) - Group discussions (see section 5.3.10) <p>Triangulation:</p> <ul style="list-style-type: none"> - Assessed artefacts and feedback (see section 5.3.6) - Observation notes (see section 5.3.7)

5.5.3 Data analysis for research question 3

In research question 3, the benefits and the challenges associated with the different assessment methods as perceived by the teachers in the study and by their students are investigated. For questions 3.1 and 3.4 (on the teachers' and the students' perception of the usability of the assessment methods), the data were analysed quantitatively using descriptive statistics. For questions 3.2 and 3.3 as well as 3.5 and 3.6 (on benefits, challenges and measures of support), respective coding systems were developed inductively and the data were analysed using content analysis (Mayring, 1994; 2010). The quality of data analysis was ensured by double-coding a portion of the data and by triangulating with other data from the study. Details will be provided below.

Procedure of quantitative data analysis for research questions 3.1 and 3.4

For research questions 3.1 and 3.4 on the teachers' and the students' perception of the usability of the assessment methods, the cases that match the criteria as specified in sub-chapter 5.4 were taken into account. The teachers' and the students' answers on respective Likert-scale items in the evaluation forms (see section 5.3.4 for teachers, 5.3.8 for students) were analysed using descriptive statistics.

Procedure of qualitative data analysis for research questions 3.2, 3.3, 3.5, and 3.6

For research questions 3.2 and 3.3 as well as 3.5 and 3.6, the cases that match the criteria as specified in sub-chapter 5.4 were analysed. The coding system was inductively developed from the data in the evaluation forms of the teachers (see section 5.3.4), in the individual interviews with the teachers (see section 5.3.9) and in the evaluation forms of the students (see section 5.3.8). The units of coding were single words. All codes were developed in German and afterwards translated. The coding system was evaluated by double-coding 18% of the teacher data (10 cases out of 57 cases) and 15% of the student data (15 cases out of 103 cases). The interrater-reliability, measured using Cohen's Kappa, was $\kappa=0.89$ for the teacher data and $\kappa=0.87$ for the student data. Landis and Koch (1977) consider values ≥ 0.81 as almost perfect agreement. The results were triangulated with the transcripts of the group discussions in the end-semester meetings with all teachers (see section 5.3.10) where the experiences with the formative assessment methods were exchanged. The triangulation showed that the coding system developed from the evaluation forms and the individual interviews was also applicable to the transcripts of the group discussions.

Whereas teachers were anticipated to perceive formative assessment as one process, the students were expected to potentially perceive peer-assessment as two processes: The first one being 'assessing peers'

and the second one being ‘receiving feedback from peers’. These potentially two processes were investigated separately in the student questionnaire (see appendix A4), and consequently, the answers were also analysed separately in the results part in chapter 7.

Coding frame for research questions 3.2, 3.3, 3.5, and 3.6

Research questions 3.2 and 3.3 as well as 3.5 and 3.6 were answered from qualitative data analysis. Table 19 provides an overview of the content of these research questions, the dimensions of the inductively developed coding frame and the data analysed for the respective research questions. The coding instructions with descriptions and examples for each category can be found in appendices A10, A11 and A12.

Table 19: Coding frame for the third research question.

Research questions	Dimensions and categories of the coding frame (highest hierarchical levels only)	Data
3.2 Benefits and challenges of different methods of formative assessment as mentioned by the teachers	- Themes emerging from the teacher’ evaluations of benefits and challenges of the formative assessment methods trialled; see appendix A10 (inductive development of codes)	- Evaluation form teachers (see section 5.3.4) - Individual interviews teachers (see section 5.3.9) Triangulation - Group discussions teachers (see section 5.3.10)
3.3 Means of support as mentioned by the teachers	- Means of support, see appendix A11 (inductive development of codes)	- Evaluation form teachers (see section 5.3.4) - Individual interviews teachers (see section 5.3.9) Triangulation - Group discussions teachers (see section 5.3.10)
3.5 Benefits and challenges of peer-assessment as mentioned by the students	- Coding frame developed from the teacher’ evaluations of benefits and challenges of the formative assessment methods trialled, see appendix A10 (inductive development of codes based on the teacher data, see 3.2)	- Evaluation form students (see section 5.3.8)
3.6 Means of support of peer-assessment as mentioned by the students	- Means of support, see appendix A12 (inductive development of codes)	- Evaluation form students (see section 5.3.8)

5.5.4 Data analysis for research question 4

In research question 4, the changes in the teachers' understanding of formative assessment and in their implementations throughout the collaboration in the study were investigated.

With the small sample sizes, the results on research question 4 are clearly tenuous. Due to the little literature on changes in teachers' formative assessment practices and beliefs throughout the collaboration in a project where the teachers develop their own assessment (see sub-chapter 3.7) available, it nevertheless appeared legitimate to conduct the respective analyses. The results will be interpreted with caution. Part of this cautious interpretation is that the data will, in some sections, not be analysed separately for the two school levels as for the other research questions. Instead, the teachers will be considered as one group.

For research question 4.1 (changes in the teachers' understanding of formative assessment), the data and the coding frame from question 1 was used but analysed dependent on the round of implementation. For question 4.2 on the formative assessment self-efficacy, data from the teacher profile questionnaire were analysed quantitatively using non-parametric tests. The data had been collected at the very beginning and at the end of the teachers' collaboration in the study. For questions 4.3 and 4.4 on the changes in the teachers' implementation and on the changes in the benefits and challenges as perceived by the teachers, the data from research questions 2 and 3.2 were now analysed dependent on the round of implementation. They were combined with quantitative data from the teacher profile questionnaire. For question 4.5 on the support mechanisms in the study, data from the individual interviews were coded inductively and analysed using content analysis (Mayring, 1994; 2010). For question 4.6 on the implementation behaviour of the individual teachers, finally, the data from research question 2 on the implementations was analysed dependent on the individual teacher. Details will be provided below.

Procedure of qualitative analysis for research questions 4.1, 4.3, 4.4, and 4.5

For question 4.1 on the changes in the teachers' understanding of formative assessment, the coding frame from research question 1 was used and now analysed taking into account the round of implementation (first, second, or third semester of collaboration in the study).

For the qualitative part of research questions 4.3 (changes in the implementations) and 4.4 (changes in the benefits and challenges perceived by the teachers), the data and the coding frames from questions 2 and 3.2 were used but analysed dependent on the round of implementation (first, second, or third semester of collaboration in the study).

For research question 4.5 (support mechanisms from the collaboration in the study), the coding system was inductively developed from individual interview data. The units of coding were single words. All codes were developed in German and afterwards translated. The data was analysed using content analysis (Mayring, 1994; 2010).

Coding frame for research questions 4.1, 4.3, 4.4, and 4.5

Research question 4.1, parts of research questions 4.3 and 4.4, as well as research question 4.5 were answered using qualitative data analysis. Table 20 provides an overview of the content of these research questions, the dimensions of the coding frame and the data analysed for the respective research questions.

Table 20: Coding frame for the fourth research question.

Research questions	Dimensions and categories of the coding frame (highest hierarchical levels only)	Coded data
4.1 Changes in the understanding of formative assessment	- Dimensions and categories from research question 1 analysed dependent on the round of implementation (first, second, and third semester of collaboration in the study)	See research question 1
4.3 Changes in the implementations	- Dimensions and categories from research question 2 analysed dependent on the round of implementation (first, second, and third semester of collaboration in the study)	See research question 2
4.4 Changes in the benefits and challenges perceived	- Dimensions and categories from research question 3.2 analysed dependent on the round of implementation (first, second, and third semester of collaboration in the study)	See research question 3.2
4.5 Support mechanisms from the collaboration in the study	- Categories on support mechanisms from the collaboration in the study (inductive development of codes; see appendix A13 for details)	- Individual interviews teachers (see section 5.3.9)

Procedure of quantitative analysis in research questions 4.2, 4.3, and 4.4

For 4.2 (changes in the self-efficacy), and for the quantitative parts of research questions 4.3 (changes in the frequency-related assessment habits), 4.4 (changes in the perception of importance of assessment), data from the teacher profile questionnaire were analysed (see section 5.3.1). For the latter two research questions, the data analysis started with the factor analysis and the formation of scales from the items included in the questionnaire. SPSS was used for this procedure. The scales for the self-efficacy were derived from the literature.

Only the data from the teachers who filled out the questionnaire at the beginning (September 2014) as well as at the end of the study (January 2016) were included. An overview of the respective data analysed can be found in Table 21.

Table 21: Teacher profile questionnaires that were included in the analysis.

	Number of teachers who filled out questionnaire at both measurement points and were therefore included in the analysis
Teachers collaborating in the study	N=16 (6 teachers from primary school, 10 teachers from upper secondary school)
Control group	N=13 (5 teachers from primary school, 8 teachers from upper secondary school)

Scales for quantitative analysis in research questions 4.2, 4.3, and 4.4

An overview of the resulting scales can be found in Table 22. The teacher profile questionnaire with all items can be found in appendix A1.

Table 22: Overview of scales built from the items of the teacher profile questionnaire.

Scale name	Items included (see appendix A1 for details)	Cronbach's α (Sept 2014 / Jan 2016)
Formative assessment – frequency (explorative factor analysis)	18a, 19a, 20a, 21a, 22a on the teachers' estimation of how often he/she uses formative assessment in her/his teaching	.78 / .78
Formative assessment – importance (explorative factor analysis)	18b, 19b, 20b, 21b, 22b, 23b, 24b, 28b, 29b on the teachers' opinion of how important distinct activities for formative assessment are	.72 / .76
Summative assessment – frequency (explorative factor analysis)	33a, 34a, 35a on the teachers' estimation of how often he/she uses summative assessment in her/his teaching	.79 / .80
Summative assessment – importance (explorative factor analysis)	33b, 34b, 35b on the teachers' opinion of how important distinct activities of summative assessment are	.02 / .71
Personal formative assessment efficacy belief (scale adapted from Enochs & Riggs, 1990)	38, 39, 40, 41, 42, 45, 47, 48, 49 adapted for formative assessment from the respective items in Enochs and Riggs (1990)	.86 / .76
Outcome expectancy (scale adapted from Enochs & Riggs, 1990)	43, 44, 46 adapted for formative assessment from the respective items in Enochs and Riggs, 1990	.01 / .45

Scales with a Cronbach's $\alpha > .7$ are typically considered acceptable (Schmitt, 1996). In this study, this is the case in both the Sept 2014 and the Jan 2016 measurement of the following scales: formative assessment – frequency; formative assessment – importance; summative assessment – frequency; personal formative assessment efficacy belief. These scales were analysed as described in the subsequent sections.

Descriptive statistics in research questions 4.2, 4.3, and 4.4

For the interval-scaled data (the scales on the importance of formative assessment and on the personal formative assessment efficacy belief), medians, arithmetic means and standard deviations were calculated for a first impression on the results. For the ordinal-scaled data (the scales on the frequency of both formative and summative assessment), medians were calculated.

Significance tests and effect sizes in research questions 4.2, 4.3, and 4.4

In order to show significant changes in the central tendency between the beginning of the collaboration in the study and its end, non-parametric Wilcoxon signed rank tests (Wilcoxon, 1945) were performed in SPSS. As the quantitative data analysed in 4.2, 4.3 and 4.4 is interval scaled, performing t-tests rather than the more conservative Wilcoxon (which involves downgrading the data to an ordinal scale) was initially considered. But Wilcoxon is robust to small samples which are not normally distributed and may contain outliers. It was therefore chosen for usage. The data from the two groups (the teachers collaborating in the study and the teachers from the control group) are analysed totally independently.

Wilcoxon signed rank tests compare, as the name indicates, signed ranks. The pairs of data points (pre and post) of all members of a group (either the teachers collaborating in the study or the teachers from the control group) are ranked. If there are changes throughout time, the resulting two total ranks will systematically differ. The Wilcoxon test statistic W is the basis of the test and it is simply the smaller of the two total ranks. For sample sizes $n > 10$ like in this study, W can be z-standardized. The z-value which will be reported in the results in sub-chapter 7.4 is

$$z = \frac{W - \mu_W}{SD_W}$$

with μ_W representing the expected value of W under null hypothesis and SD_W representing the standard deviation of W .

μ_W is calculated as

$$\mu_W = \frac{n_{\text{red}}(n_{\text{red}} + 1)}{4}$$

With n_{red} representing the number of pairs of data points which have a difference $\neq 0$.

This z-value can be tested for significance by comparing it to the critical value of a standard normal distribution (reported in respective tables). If the value of the test statistics is higher than the critical value, the difference is significant. This significance is reported as p in the results in sub-chapter 7.4.

For this study, the problem with significance as a measure of change is that it is sensitive to the sample size: It is difficult to measure significant changes with small sample sizes. Therefore, effect sizes which are independent of the size of the sample were also calculated. The effect size is a measure for the strength of a phenomenon such as a difference, change, or correlation between two variables. Cohen's d (Cohen, 1988) is a measure of effect size which is based on the differences between the two arithmetic means, in the case of this study the difference between the mean value of a group at time T1 and the respective value at time T2.

$$d = \frac{AM_{T1} - AM_{T2}}{\sqrt{\frac{SD_{T1} + SD_{T2}}{2}}}$$

with AM=arithmetic mean and SD=standard deviation. The effect size Cohen's d is reported in the results in sub-chapter 7.4.

Procedure of quantitative analysis in research question 4.6

For research question 4.6 (variability of implementations within teachers), the data on the trials from research question 2 were analysed on the basis of the individual teacher. For the analysis, new variables, called 'overlaps', were defined in the different sub-categories from research question 2.1 and 2.2 (such as dimensions of openness, inquiry activities etc.).

'Overlap' in the context of the different variables (such as dimensions of openness, inquiry activities etc.) means the size of the intersecting set of options in relation to the total size of options of the same variable. If, for example, a teacher's first trial was coded as open in dimensions of openness A, B, and C, and the same teacher's second trial was coded open in dimensions of openness C and E, the overlap is 0.25.

The nine newly defined variables are:

- 01 = Overlap of trials in dimensions of openness
- 02 = Overlap of trials in inquiry activities
- 03 = Overlap of trials in competences assessed
- 04 = Overlap of trials in communication of criteria
- 05 = Overlap of trials in sources of data
- 06 = Overlap of trials in assessment methods
- 07 = Overlap of trials in engagement with feedback

$O8$ = Overlap of trials in cycle length

Sum = Sum of $O1$, $O2$, $O3$, $O4$, $O5$, $O6$, $O7$, and $O8$

The variables $O1$ to $O8$ can vary between 0 and 1; the variable Sum can vary between 0 and 8.

For this, only the teachers who had completed and sufficiently documented at least two subsequent trials were included. This means that trials which were not conducted in the context of inquiry (criterion 3 in chapter 5.4) or where the formative assessment was not completely successful (criterion 4 in chapter 5.4) were still included.

Descriptive statistics in research question 4.6

Medians, arithmetic means and standard deviations were calculated for a first impression of the results.

Significance tests and effect sizes in research questions 4.6

In order to test for significant differences in the central tendency between the different subgroups of the teachers collaborating in the study, non-parametric Mann-Whitney-U tests (Mann & Whitney, 1947) were performed in SPSS. Potential subgroups were formed according to the demographic variables as specified in sub-chapter 5.2: The school level (primary school vs. upper secondary school); the gender (male vs. female) and the teaching experience (because of the small sample sizes, only two sub-groups were formed here: teaching experience 0-10 years vs. teaching experience >10 years).

Mann-Whitney-U tests are suitable to test the null hypothesis that two samples come from the same population against an alternative hypothesis, namely that there are two significantly different sub-samples. As the quantitative data analysed in 4.6 is interval scaled, performing t-tests rather than the more conservative Mann-Whitney-U tests (which involves downgrading the data to an ordinal scale) was initially considered. But the U-test is robust to small samples which are not normally distributed and may contain outliers. It was therefore chosen for usage.

Similar to Wilcoxon signed rank tests introduced above, Mann-Whitney-U tests compare ranks. The data from the whole group are ordered by size and ranked starting with 1 for the lowest value. All ranks of the two subgroups suspected (such as primary school and upper secondary school teachers) are then added up to two sums, called total ranks. If there really are two subgroups in the whole group, the two total ranks will systematically differ. The Mann-Whitney statistic U is the basis of the test and it is simply the higher of the two total ranks. The *U-value* is reported in the results in sub-chapter 7.4. For sample sizes $n < 20$ like in this study, U cannot be z-standardized. This means that the sample is too small to compare the U-value to the critical value of a standard normal distribution and to report the significance as *p*. SPSS automatically adjusts the test respectively and calculates so-called exact significances which are reported as *p* ($2 * (1 \text{ tailed})$) in the results in section 7.4.6.

For this study, the problem with significance as a measure of difference is that it is sensitive to the sample size: It is difficult to measure significant differences with small sample sizes. Therefore, effect sizes which are independent of the size of the sample were also calculated. The effect size is a measure for the strength of a phenomenon such as a difference, change, or correlation between two variables. Cohen's *d* (Cohen, 1988) is a measure of effect size based on the differences between the two arithmetic means; in this case the difference between the mean value of a subgroup 1 and the respective value of a sub-group 2.

$$d = \frac{AM_1 - AM_2}{\sqrt{\frac{SD_1 + SD_2}{2}}}$$

with AM=arithmetic mean and SD=standard deviation. The effect size Cohen's *d* is reported in the results in section 7.4.6.

6 Illustrative examples of implementations

In this chapter, three examples of implementations will be introduced, one for each assessment method. The aim is to provide an idea of what the trials looked like with concrete cases and to exemplify the nature of data that will later be analysed.

6.1 Written teacher assessment at primary school

A teacher at 3rd grade primary school level (P5) let her students observe the growth of chicks. So she had a group of chicks in her classroom for almost a month and the students observed them every morning for 5-10min. For this, the students had to think of specific questions, to focus their observations on these questions, and they also had to individually write a short paragraph on their observations in the so-called chick journal every morning. The teacher collected the chick journals every day and shortly commented on the latest student entries (see Figure 7). She focussed her feedback on (a) the distinction between observations and claims/conclusions and (b) on the precision of the descriptions and sketches. The students were often told to read the teacher's comments and to check for improvement but also to read the journal entries of their peers. In addition to this, the teacher planned and conducted several short inputs and group discussions on good observations and on the distinction between observations and claims throughout the unit.

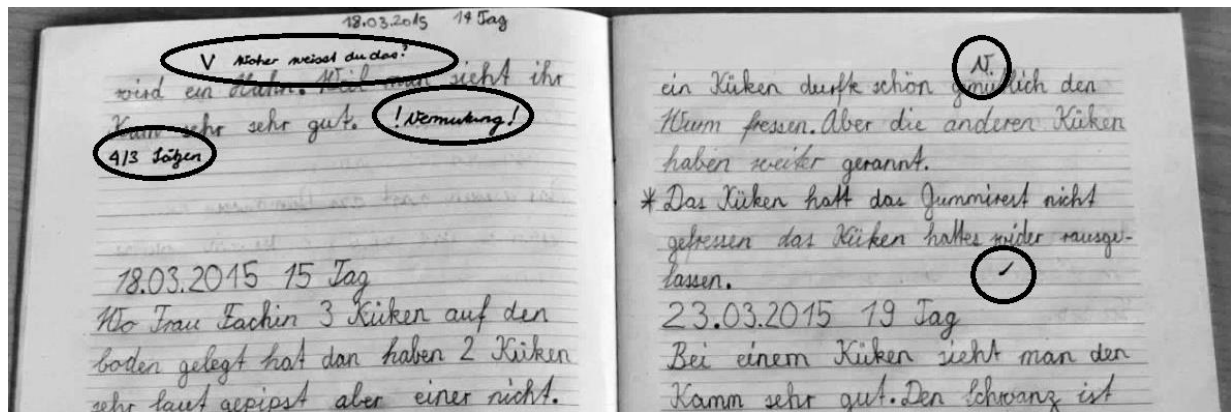


Figure 7: Example of a student's chick journal with the teacher's comments (teacher's comments circled in black).

In the evaluation form, the teacher reported that "pointing at these two assessment criteria over a long period of time really helped the students to keep focussed". She said in the interview that "it was easy to remind both the students but also their parents of the goals of that unit". The teacher also realized that "it took some time before the students really understood the assessment criteria but also the principle of this type of formative feedback. Some of the students really needed the external encouragement to engage with the feedback and to plan next steps in learning". She also mentioned that "the extra workload for providing feedback to every student every day is only worth the effort when the assessment criteria are very relevant in science education". So she recommended "spending ample time of the preparation phase on the selection of suitable criteria".

In the group discussion where the teacher told about her experiences with this trial, her teacher colleagues were immediately convinced about the effect of the many rounds of individual feedback and the many subsequent opportunities for the students to engage with this feedback. However, they shared the teacher's impression that such a time-consuming, demanding setting needs careful placement in the semester and thoughtful selection of the assessment criteria.

The codes ascribed to this trial can be found in appendix A7 (case from teacher P5 in the second round of implementation).

6.2 Peer-assessment at upper secondary school

In upper secondary physics education, a teacher (S1) set up learning stations on stationary waves including, for example, producing sound with rulers, bottles and glasses or Kundt dust patterns. The instructions provided little guidance and encouraged the students to explore. The students worked in groups, each group summarized their observations and conclusions from the different learning stations in a report. Afterwards, the student groups exchanged their reports and provided each other with feedback, scaffolded by the structuring aid displayed in Figure 8. All student groups then got their own reports together with feedback from peers back. They had time to discuss the feedback received and to decide if and how to improve the report. The final report was then graded by the teacher.

- Advisement for revision of the reports**

 - 1) Is the report complete? Is it obvious which parts belong to which learning station?
 - 2) Are all answers being answered?
 - 3) Are the explanations complete?
 - 4) Are the explanations understandable and logical?
 - 5) Are there any contradictions in the explanations?
 - 6) Are there any graphics to clarify issues? Would graphical representations help?
 - 7) What is missing so that I would understand the explanations?
 - 8) What would I have done differently?
 - 9) What amendments would I make?
 - 10) What is good?

Figure 8: Structuring aid for peer-assessment as developed by a teacher.

The assessment of the reports by peers triggered intense discussions and explanations on the content. In the individual interview, the teacher put it like this: *“Many students were unsure whether they understood their peers’ reports and directly approached other groups in order to ask what was meant by a paragraph of writing or similar. This is a nice side-effect, the students started to talk about physics and not only about sneakers and TV series and mobile phones.”* When it comes to drawbacks, the teacher stressed that *„the students felt unsure about assessing each other, whether the comments for the peers were correct.”* Another issue was that some student groups took the provision of advice for peers very serious whereas other student groups did not. Thinking about her next trial in the subsequent semester, the teacher said: *“I might probably try to come up with some incentive that prompts all students to take their job serious. Otherwise, it is not fair. Some students write careful pieces of advice and receive a botch in turn.”* Overall, the teacher had a positive impression of the peer-assessment; she summarized the trial as follows: *“Yes, the peer-assessment was worth a trial. The students enjoyed it and they also told me that assessing peers was something different and interesting. Since the students asked questions to each other and discussed problems among them, I did not have to take this responsibility but had the opportunity to observe and overhear discussions and get a more detailed insight into the students’ thinking.”*

This was one of the trials where the students’ perspective was captured by a questionnaire with open-ended items. The students reported that they learned in terms of different dimensions: They were able to improve their reports through detailed suggestions from the feedback. But the peer-assessment also provided the students with the possibility to look at their peers’ work and therefore to develop a broader horizon of possible solutions. Addressing the problems and challenges associated with peer-assessment, the two main issues also brought up by the teacher were mentioned: Firstly, some students felt unsure whether their feedback was valid and whether they would be able to complete a “teacher duty”. Secondly, it was criticized that the peer-assessment was only valuable when the student(s) providing the feedback engaged seriously and carefully with the draft reports.

The codes ascribed to this trial can be found in appendix A7 (case from teacher S1 in the first round of implementation).

6.3 Combination of peer-assessment and self-assessment at primary school

In a mixed class of 3rd and 4th year primary school students, the teacher (P9) organized half-day excursions to the local forest every month. In every excursion, the students explore an aspect related to this ecosystem. This particular excursion was aimed at investigating soil profiles in the forest. Before leaving the classroom, the investigation was planned: Tools to dig in the ground had to be organized, student groups had to be built and suitable questions to guide the investigations had to be developed.

In the forest, the students were expected to dig holes to explore the soil at different places. The students conducted their investigative work on the holes: They measured the depth of the different layers, described their appearance and the abundance of leaves and animals, took samples of the different layers and made sketches.

Back in the classroom, every group created a poster with their findings. But before starting the work on the posters, the students reflected on the quality of their investigations (documentation of findings on place with written notes and sketches), on their attitude to work (precise and exact style of conducting the investigation), and on their collaboration in the group (engagement in the discussions in the group). The self-assessment was guided by the questions on the upper part of the worksheet displayed in Figure 9 and by an estimate of their own portion of the group effort. After reflecting upon their own work, the students had to provide feedback to the peers in their group on their engagement. This is displayed in the lower part of the worksheet in Figure 9. Parts of the formative assessment activities were already known to the students from an earlier sequence.

Since the students visit the forest every month and explore a particular topic in a group with other students, the teacher expected his students to use the reflections and the peer-feedback in the later visits.

Ich habe zuerst Fragen zum Boden untersuchen aufgeschrieben
trifft zu <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> trifft nicht zu
Ich habe mir Gedanken gemacht, wie ich mit der Gruppe arbeiten will.
trifft zu <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> trifft nicht zu
Ich habe eine Skizze vom Boden gemacht.
trifft zu <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> trifft nicht zu
Die Skizze ist beschriftet.
trifft zu <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> trifft nicht zu
Ich habe in der Gruppe über die Fragen nachgedacht.
trifft zu <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> trifft nicht zu
Ich habe sorgfältig und genau gearbeitet
trifft zu <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> trifft nicht zu
Mein Anteil an der Gruppe: _____ (alle 4 zusammen = 100)
Dies sage ich meinen Gruppenmitgliedern:

Figure 9: Form for self-assessment (upper part) and peer-assessment (lower part) on the exploration of a soil profile and the work in the student group.

In the individual interview, the teacher spoke about how he planned this unit in the forest and the formative assessment: „In terms of lesson preparation, I am somewhat chaotic. I start with the topic [...], and the assessment is something that arises naturally. [...] When I notice that a particular formative assessment activity would just fit.”

On this particular worksheet for self- and peer-assessment, the teacher mentioned that his students were generally familiar with the reflective questions after the excursions to the forest, even to the particular task where the own share of the total group effort is estimated. Looking at the upper part of his worksheet again, the teacher said: “[...] I am not sure I formulated these questions careful enough. For some of the answers, it is hard to provide evidence, the students' self-assessment cannot really be proven. Like, 'I gave some thought on

how to support the group', how can anybody control that? [...] These questions are not very good, they should be much clearer, like P2 did." P2 was another primary school teacher who also used prompting questions for peer-assessment. The teachers knew about each other's trials at that time because the end-of-semester meeting had just taken place where all teachers exchanged their experiences on what they had done.

Speaking about the value of formative assessment, P9 said: *"In science education and particularly in such open settings, I make only few classical exams on content knowledge. Instead, I am continuously searching for possibilities and tailoring units where I can obtain the information necessary for the grades, like the poster in this example, and at the same time support the students holistically. The students appreciate these, sometimes fancy, approaches to developing their personalities."*

The codes ascribed to this trial can be found in appendix A7 (case from teacher P9 in the first round of implementation).

7 Results

The results chapter is structured in four parts following the four research questions. In the first sub-chapter, research question 1: *'The teachers' understanding of formative assessment'* is analysed. The data includes the teachers' written definitions of formative assessment which are analysed qualitatively.

The second sub-chapter focusses on research question 2: *'How do the teachers in the study trial formative assessment methods in their inquiry teaching?'* Within this sub-chapter, the different trials will be characterized and analysed in terms of their inquiry nature (section 7.2.1), in terms of the formative assessment methods (section 7.2.2) and in terms of problems that occurred (section 7.2.3). Research question 2 is investigated by analysing the teaching plans from the teacher evaluation forms and the associated teaching materials as well as the individual interviews and the group discussions. The analysis will be done separately for the cases that match the criteria as specified in sub-chapter 5.4 and the cases that did not match those criteria.

In the third sub-chapter, the data relating to research question 3: *'How do the teachers and the students evaluate the formative assessment methods trialled?'* will be presented. Consequently, the aim of this sub-chapter is not to know what the teachers did (as in the second sub-chapter) but to reconstruct what both the teachers and the students thought about the methods trialled. In the first section of the sub-chapter, the teachers' perceptions of benefits and challenges related to the different assessment methods which they have trialled will be presented (sections 7.3.1 and 7.3.2). This is investigated by analysing the teacher evaluation forms as well as individual teacher interview data. In the later sections of sub-chapter 7.3, the evaluation form for students will be analysed (sections 7.3.4 and 7.3.5). Both the teachers and the students were asked to suggest possible means of support that might facilitate the formative assessment practices. The respective answers will be enclosed in sections 7.3.3 and 7.3.6.

In the fourth sub-chapter, research question 4: *'Changes in the teachers' understanding and implementation of formative assessment throughout the study'* will be investigated. The aim of this sub-chapter is to search for possible changes in the teachers' understanding of formative assessment (section 7.4.1); changes in their self-efficacy (section 7.4.2); changes in the teachers' formative assessment practices (section 7.4.3), changes in their perceptions of importance, benefits and challenges related to formative assessment (section 7.4.4); and the support mechanisms from the collaboration in the study as mentioned by the teachers (section 7.4.5). The last section (7.4.6) summarizes the teachers' implementation behaviours as represented in the variability of their trials. The data is the same which had already been coded for research questions 1, 2, and 3, but is this time analysed dependent on the round of implementation. Additionally, the teacher profile questionnaire will be analysed using non-parametric tests.

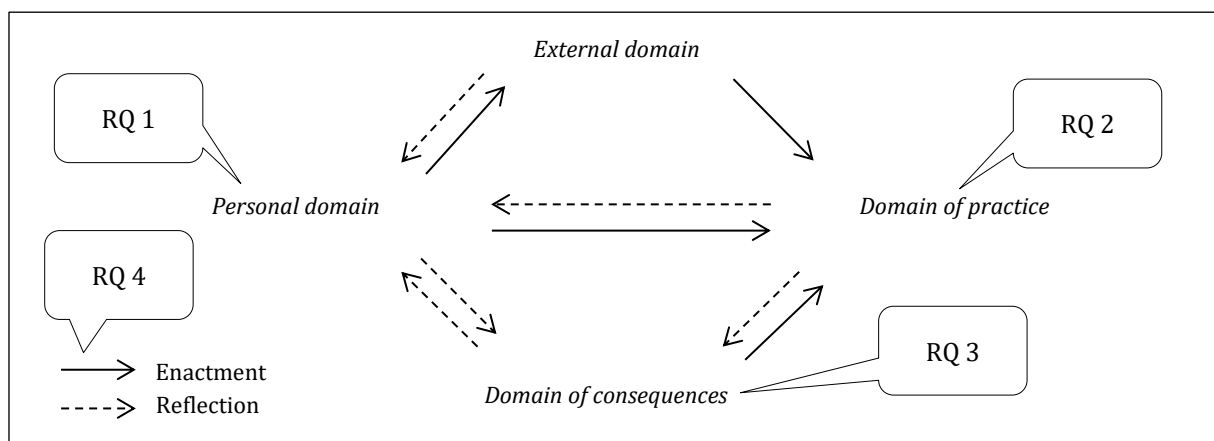


Figure 10: Relations between the model of professional growth (Clarke & Hollingsworth, 2002), and the results on the four research questions in this study.

As pointed out in sub-chapter 4.1, the four research questions and the corresponding results can, to some extent, be related to the domains in the model of professional growth (Clarke & Hollingsworth, 2002). The model was developed to provide a theoretical frame for innovation in teaching. As visualised in Figure 10, the results to research question 1 can be related to aspects of the personal domain. The results to research question 2 target the professional experimentation in the domain of practice. The results to research question 3 can be interpreted as insights to the domain of consequences. Finally, the results to research question 4 will explore the interdependencies between the different domains.

7.1 Results on research question 1: Teachers' understanding of formative assessment

In this sub-chapter, research question 1: '*The teachers' understanding of formative assessment*' is analysed. The data includes the written definition of formative assessment from teachers which is analysed qualitatively.

The following Table 23 displays the 11 different elements that were mentioned by the teachers in their written explications of what formative assessment is. These definitions were collected three times throughout the study (at the very beginning, in the middle, at the very end). Most of the definitions contained more than one element which is why the total number of elements mentioned is much higher than the number of teachers. The elements mentioned in the definitions were organized into four groups.

Table 23: Teachers' understanding of formative assessment.

		Number of definitions referring to the element at primary school (n=22 definitions)	Number of definitions referring to the element at upper secondary school (n=29 definitions)
Elements that are ascribed to formative assessment in the literature, too	a) Supportive in nature	9	9
	b) Providing guidance on next steps in learning for students	8	9
	c) Providing guidance on next steps in teaching for teacher	4	9
	d) Individual and/or part of differentiation	2	1
	e) Prospective rather than retrospective in nature; opposite to summative assessment	10	10
Elements that are ascribed to assessment in general (not specific for formative assessment, though)	f) Criterion-based	3	3
Other elements	g) Focussed on a specific set of competences or other learning goals	0	3
	h) Having an individual reference norm	1	0
	i) Grading of the learning process	2	2
	j) Unclear/ reference to inquiry features	7	8
Examples of formative assessment methods from the study	k) Examples of assessment methods	4	6

These elements will be introduced in more detail in the following.

Supportive in nature

This category was used to code teacher quotes that conveyed the idea that formative assessment was supporting the students or the students' learning. Illustrative quotes (translated from German; not allocated to

a particular teacher because of the anonymous data collection) include: *“formative assessment is meant to support the learning of the students”*; *“supportive feedback”*.

Providing guidance on next steps in learning for students

This category summarizes the idea that the main aim of formative assessment is to provide guidance on the next steps in learning for the students. Illustrative quotes (translated from German; not allocated to a particular teacher because of the anonymous data collection) include: *“Where am I, where do I want to go”*; *“describes what students already know and also describes what they do not know yet”*; *“provide guidance”*; *“coach students in their work”*; *“hints without giving away the solution”*.

Providing guidance on next steps in teaching for teacher

This category encloses quotes from teachers who described the main aim of formative assessment as providing guidance on next steps in teaching for the teacher (rather than next steps in learning for the student as in the last category). Illustrative quotes (translated from German; not allocated to a particular teacher because of the anonymous data collection) include: *“I see what is there (preconcepts) and build my teaching on it”*; *“based on these insights I can tailor my teaching”*; *“depending on these “checkpoints” I can intervene”*.

Individual and/or part of differentiation

This category was used to code teacher quotes that explained that formative assessment was individual assessment (rather than an overview-type of assessment of the whole classroom) or that formative assessment was part of differentiation. Illustrative quotes (translated from German; not allocated to a particular teacher because of the anonymous data collection) include: *“Individual assessment for each student”*; *“individual progress”*; *“shows where the individual problems are”*; *“insight to a person’s actual level of performance”*; *“individual, so different for different students”*; *“differentiation dependent on what each student already knows”*.

Prospective rather than retrospective in nature; opposite to summative assessment

This category encloses quotes that described formative assessment as prospective rather than retrospective, meaning that it was focussed on future learning rather than making up the balance of what was achieved. Closely related to that understanding and therefore included in the same category was the description of formative assessment as a kind of counterpart to summative assessment. Illustrative quotes (translated from German; not allocated to a particular teacher because of the anonymous data collection) include: *“Does not involve grading”*; *“formative assessment is a counterpart to summative assessment”*; *“it is about the future learning”*; *“formative assessment is provided before the end of the learning process; so during the learning process”*.

Criterion-based

This category summarizes quotes which conveyed the idea that formative assessment was criterion-based. Illustrative quotes (translated from German; not allocated to a particular teacher because of the anonymous data collection) include: *“Teacher observes and takes notes in a rubric on what can be observed”*; *“criteria are pre-defined”*.

Focussed on a specific set of competences or other learning goals

This category was used to code teacher quotes which conveyed the idea that formative assessment describes assessment of a specific set of competences or learning goals, such as social competences. Illustrative quotes (translated from German; not allocated to a particular teacher because of the anonymous data collection) include: *“Feedback on the personality of the student”*, *“assessment of the students’ self-regulation skills”*; *“focussing on social competences”*.

Having an individual reference norm

This category was used to code teacher quotes which conveyed the idea that formative assessment had an individual reference norm. Illustrative quotes (translated from German; not allocated to a particular teacher

because of the anonymous data collection) include: *“In formative assessment, the students is assessed based on individual learning goals”*; *“assessment based on the individual progress”*.

Grading of the learning process

This category was used to code teacher quotes which conveyed the idea that formative assessment meant grading of the learning progress rather than grading of the product. Illustrative quotes (translated from German; not allocated to a particular teacher because of the anonymous data collection) include: *“Assessment which does not focus on the product but on the learning process”*.

Unclear/ reference to inquiry features

This category summarizes teacher quotes that were unclear or that referred to inquiry features. Illustrative quotes (translated from German; not allocated to a particular teacher because of the anonymous data collection) include: *“Organise the group, discuss the plans, distribute task amongst group members, documents problem solving process”*; *“the assessment should provide comparable results to students”*.

Examples of assessment methods

In this category, references to a particular assessment method were summarized. Illustrative quotes (translated from German; not allocated to a particular teacher because of the anonymous data collection) include: *“Written teacher feedback”*; *“peer-assessment”*; *“self-assessment”*.

Distribution of categories at the two school levels explored

The majority of categories were covered by quotes of teachers from both school levels. This is not the case for two categories: Element g) conveying the idea that formative assessment focusses on a specific set of competences or other learning goals was only mentioned by upper secondary school teachers. Element h), which describes formative assessment as having an individual reference norm, was only mentioned at primary school level.

The teachers' descriptions of the term formative assessment were analysed. The teachers were asked to write down an explanation of the expression three times during the study (beginning of collaboration; middle phase; end of collaboration). From all the answers, 11 categories were developed: Five categories that describe elements also present in the literature on formative assessment (supportive nature; guidance about next steps in learning or in teaching; individual; prospective nature); one element that is generally described to assessment in the literature (criterion-based nature); four categories with other elements (focused on a specific set of competences such as behaviour in groups; individual reference norm; grading of the learning process; unclear or reference to inquiry) and one category with examples of assessment methods.

7.2 Results on research question 2: Description and analysis of the teachers' trials

7.2.1 Description of the inquiry units in the trials

In this sub-chapter, the inquiry units used for trialling formative assessment methods will be characterized. The cases that matched the criteria from sub-chapter 5.4 will be included in this part of the results.

Dimensions of openness in the inquiry units

The dimensions of openness in the inquiry units as trialled by the primary and the upper secondary school teachers were classified according to their openness (Priemer, 2011; see sub-chapter 2.1).

The results are displayed in Figure 11. In total, 34 cases were classified (14 primary, 20 upper secondary school cases; see chapter 5.4). Out of these 34 cases, 29 cases were classified as being open in more than one dimension. This is why the numbers in Figure 11 add up to much more than 34.

Overall, the trend of having the content and the strategy of the inquiries pre-defined but giving the students some freedom in deciding about the solution processes occurs at both school levels. In about half of the inquiries at both school levels, more than one solution was possible (openness in the number of solutions). Into this group falls the first illustrative example from chapter 6, the observation of chicks: The teacher pre-defined that the topic of the investigation should be the behaviour of living chicks and that respective data should be collected by the observation of a group of chicks in the classroom. The students collaborated in groups of 3 to 4 people and every group focussed on a different aspect of the behaviour of the chicks. In terms of solutions and solution processes, the inquiry was open: The student groups decided themselves whether to discuss their observations before starting to take down their notes or to write their observational notes first and revise them together. Apparently, a variety of solutions was correct.

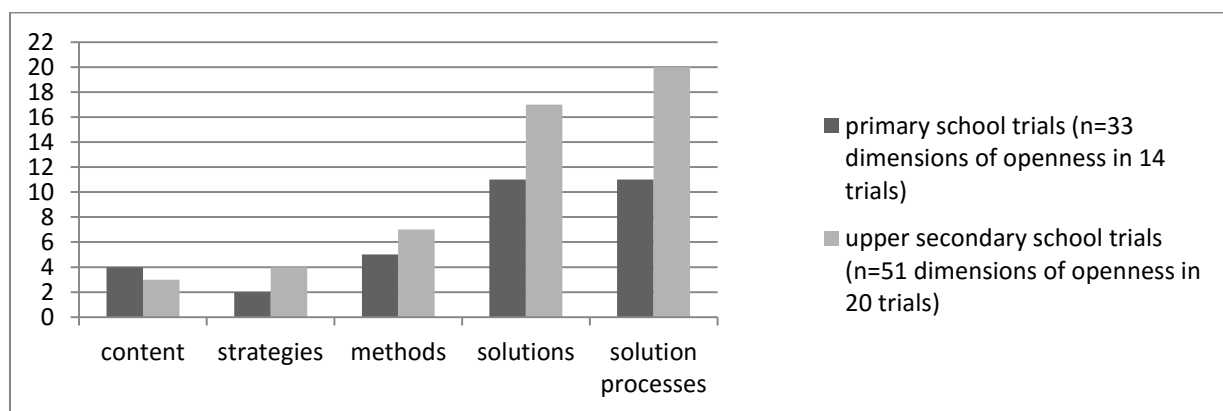


Figure 11: Dimensions of openness (Priemer, 2011) in the units of the primary and the upper secondary school teachers.

Looking at the dimensions of openness per trial in more detail, Figure 12 reveals that at primary school level, most trials were open in three dimensions whereas at upper secondary school, most trials were open in two dimensions (number of solutions and number of solution processes in all cases). Three trials at upper secondary school were open in five dimensions. All these three trials took place in the context of *Mini-Maturaarbeiten* (project to prepare for the matura thesis) or *Maturaarbeiten* (matura thesis).

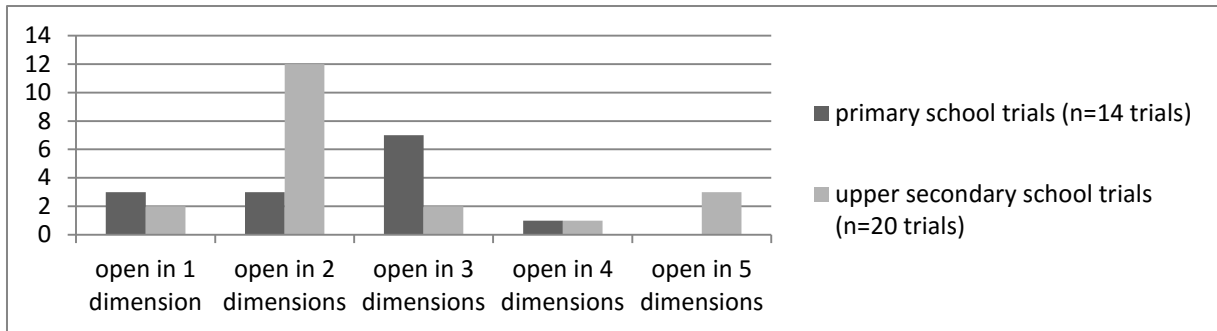


Figure 12: Number of open dimensions (Priemer, 2011) per trial in the units of the primary and the upper secondary school teachers.

Inquiry activities enacted in units

In the next step of analysis, the students' inquiry activities were categorized using Bell et al. (2010) as a theoretical basis (see sub-chapter 3.1). Figure 13 shows that the activities occurring most often were 'planning'; 'investigation'; 'analysis and interpretation'; and 'communication' at both school levels. 'Conclusion and evaluation' occurred frequently in the upper secondary school cases but not at primary school level. Other activities were performed more seldom at both school levels: 'Orienting and asking questions'; 'hypothesis generation'; 'model'; and 'prediction'.

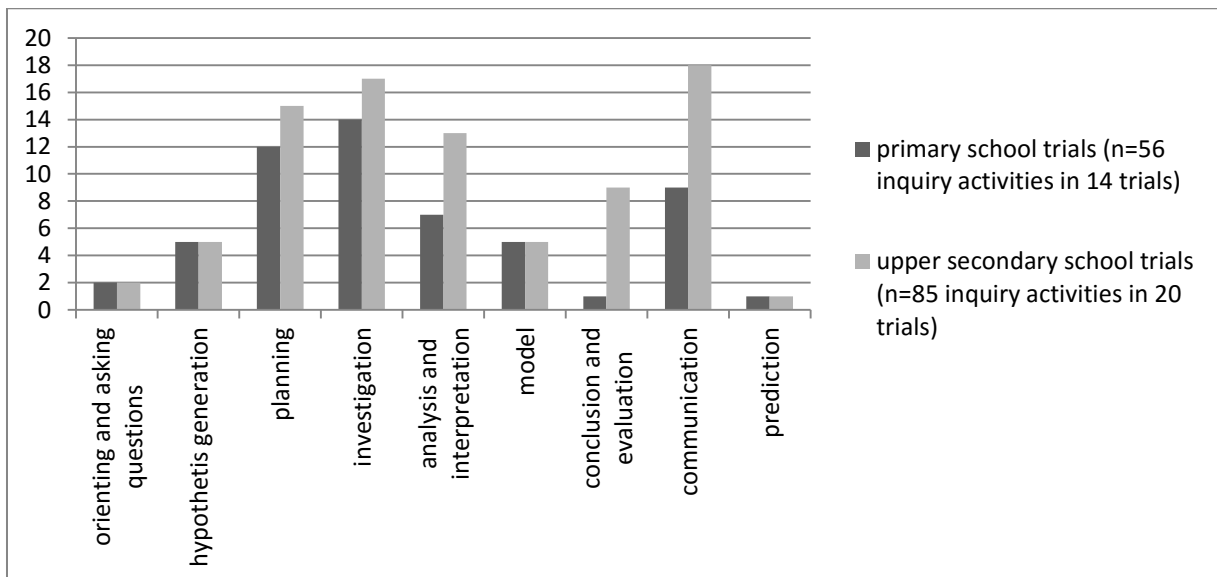


Figure 13: Inquiry activities (Bell et al., 2010) enacted in the units at primary and at upper secondary school.

Looking into the inquiry activities performed in the trials in more detail, Figure 14 shows the number of activities enacted per trial: At primary school level, the trials varied between two and six activities. Most trials involved four inquiry activities. At upper secondary school, the results varied between one and seven activities per trial. Most trials included between four to six inquiry activities. The two upper secondary school cases that included seven inquiry activities took place in the context of *Mini-Maturaarbeiten* (project to prepare for the matura thesis) or *Maturaarbeiten* (matura thesis) as mentioned earlier.

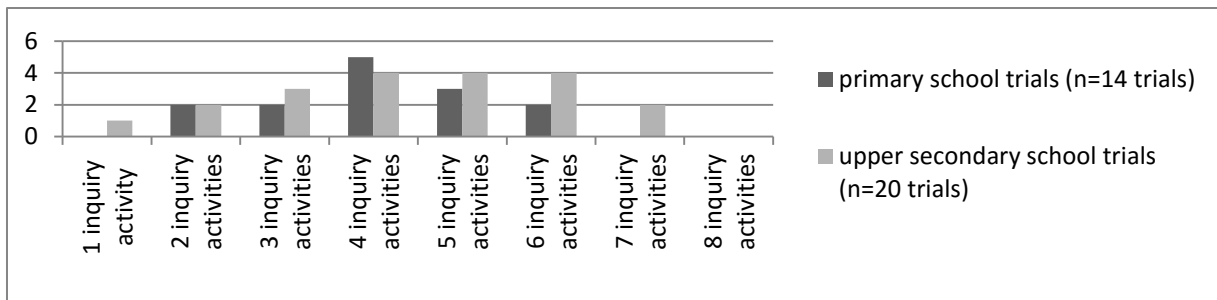


Figure 14: Number of inquiry activities per trial at primary and at upper secondary school.

Competences assessed

As laid out in section 3.1.2, a number of domain-specific and transversal competences are ascribed to inquiry. The teachers' trials were analysed in terms of what competences they assessed. The results can be found in Figure 15 (domain-specific competences) and Figure 17 (transversal competences). In total, 34 trials were classified (14 primary, 20 upper secondary school trials). In many of the trials, more than one competence was trialled (see Figure 16). This is why the numbers in the figures add up to much more than 34. In addition to the domain-specific competences, criteria covering formal aspects and content were often assessed at both school levels. They are not covered in the analysis.

Almost all domain-specific competences as defined in Bell et al. (2010) were assessed in at least one trial in the project. However, the teachers from both school levels tended to assess the 'investigation' as well as the 'communication' competences most frequently. The later includes the documentation in lab journals as well as the presentation of results in short talks and similar. At upper secondary school, 'planning' as well as 'analysis and interpretation' were also assessed rather often. The three illustrating examples in chapter 6 are in that sense typical for the trials in the study: In the first and in the second illustrating example, the documentation of the respective inquiries was assessed, in the last illustrating example on the soil profiles, the investigation was assessed.

Other domain-specific competences were rarely or never assessed at the two school levels explored: 'Orienting and asking questions', 'hypothesis generation', 'model', 'conclusion and evaluation', and 'prediction'. Looking at the similarities and differences between the two school levels, it appears that at primary school level, 'orienting and asking questions', 'model', 'conclusion and evaluation', as well as 'prediction' was not assessed at all, resulting in a narrower range of coverage compared to the situation at upper secondary school.

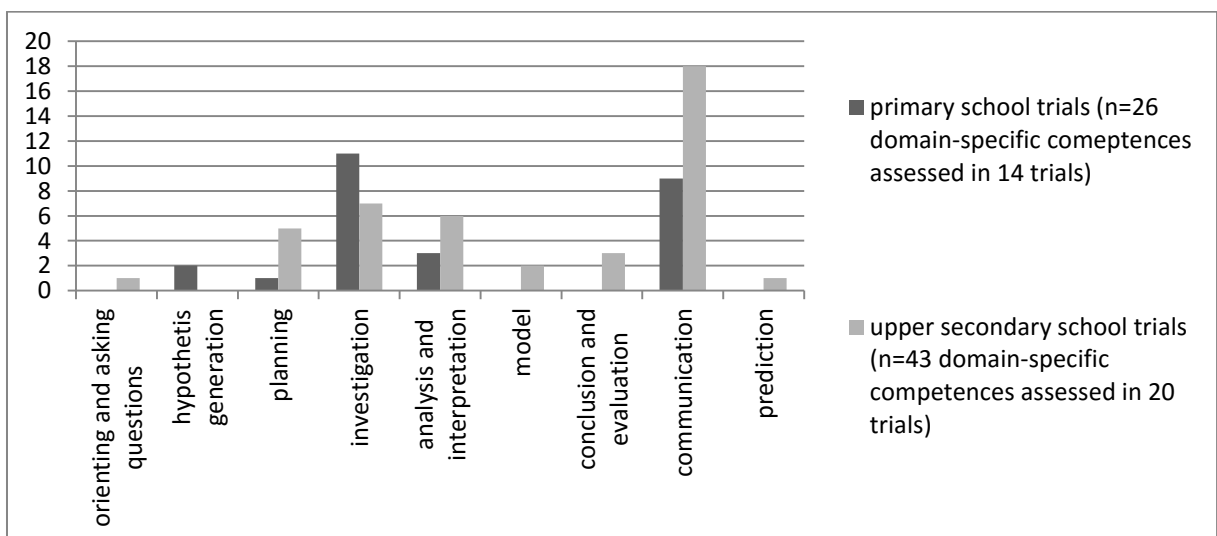


Figure 15: Domain-specific competences (see section 3.1.2) assessed.

Looking at the individual trials in more detail, the number of domain-specific competences assessed in every trial was analysed. The results are displayed in Figure 16. It can be seen that most trials at both primary and upper secondary school levels involved the formative assessment of one or two domain-specific competences. The three upper secondary school cases where five or six domain-specific competences were analysed took place in the context of *Mini-Maturaarbeiten* (project to prepare for the matura thesis) or *Maturaarbeiten* (matura thesis) as mentioned earlier.

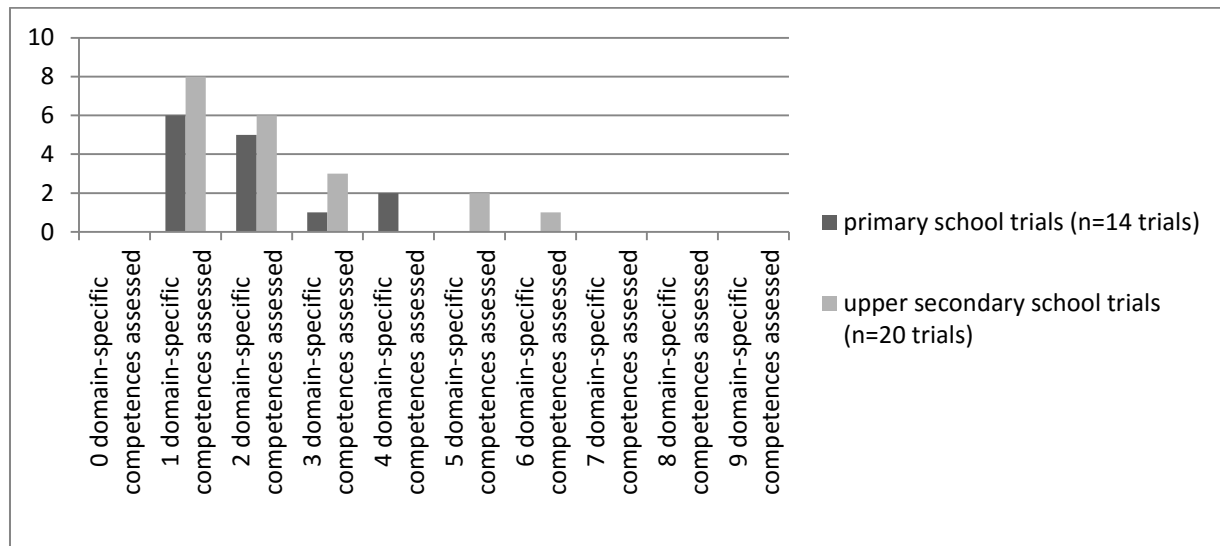


Figure 16: Number of domain-specific competences assessed per trial at primary and at upper secondary school level.

In the individual interviews, the teachers were asked about the reasons for the domain-specific competences chosen for assessment. At the primary school level, there were three basic lines of argumentation:

- No explicit decision but *“organic growth”* (teacher P9); *“suitable”* (teacher P2); *“it just emerged”* (teacher P6)
- Resource-based decision: *“I found the rubric and thought it was good”* (teacher P4)
- Decision related to the general relevance of the competence: *“Important in science education”* (referring to ‘investigation’; teacher P5)

At upper secondary school level, three lines of argumentation were found, too. There is some overlap with the primary school teachers:

- Decision based on the students’ abilities: *“had the impression that the students would be able to assess this”* (referring to ‘communication’; teacher S1)
- Decision related to the relevance of the competence in the students’ further education: *“appeared important in order to be prepared for university”*; *“relevant for the Matura thesis”* (both quotes referring to ‘communication’; teachers S2; S11)
- Decision related to the general relevance of the competence: *“important in science education”* (referring to ‘planning’ in case of teacher S2; to ‘modelling’ in case of teacher S10; to ‘conclusion and evaluation’ in case of teacher S4; to ‘communication’ in case of teacher S7).

Apart from the domain-specific competences, transversal competences were also assessed in the trials of the study. All transversal competences as defined in OECD (2015b) were assessed in several trials at both school levels explored in the study (see Figure 17). An example is provided in sub-chapter 6.3 where the students self-and peer-assessed their interaction in heterogeneous groups and their acting autonomously in an investigation on soil profiles.

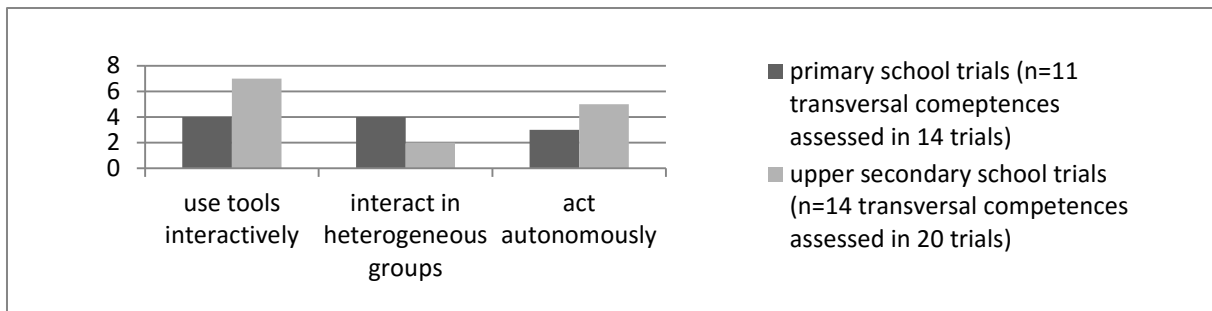


Figure 17: Transversal competences (see section 3.1.2) assessed in the study.

The transversal competences were never assessed exclusively but always in combination with one or several domain-specific competences in the same trial. However, in some of these cases, transversal competences were covered by one assessment method and domain-specific competences were covered by another assessment method in the same trial. An example is the first trial of primary school teacher P3 (see appendix A7 for details): In a unit where the students constructed a model to explain astronomical phenomena, the teacher focused her assessment on the modelling process whereas the student peers assessed each other's interaction in the student group.

Exploring the assessment of the transversal competences in the study in more detail, the number of competences assessed in every trial was displayed in Figure 18. It can be seen that between zero and two transversal competences were assessed per trial at both school levels involved in the study. Most trials at primary school level included 1 transversal competence for assessment, whereas most upper secondary school did not involve assessment of any transversal competences.

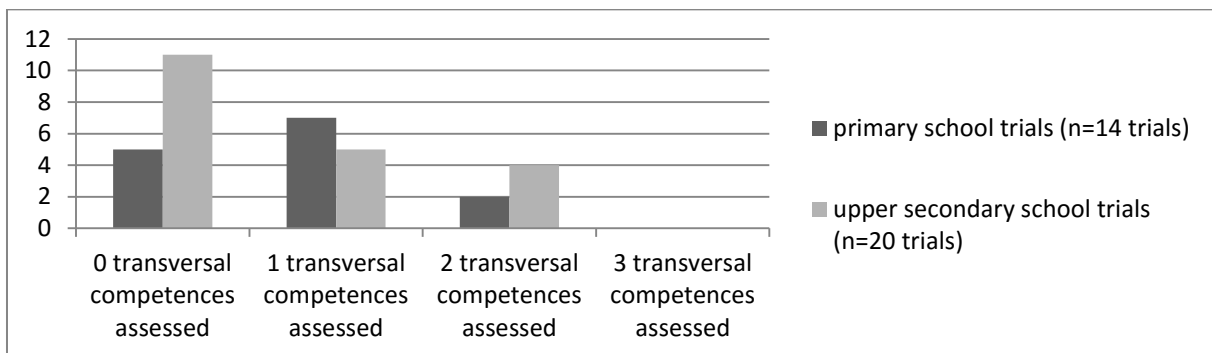


Figure 18: Number of transversal competences assessed per trial at primary and at upper secondary school level.

In the individual interviews, the teachers were asked about the reasons for the choice of a particular transversal competence for assessment. At the primary school level, only one teacher who covered transversal competences in his trial was interviewed. Therefore, there is only one line of argumentation:

- Decision related to the relevance of the competence for the personal development of the students: *"Formative assessment provides the opportunity to support the students in their development towards responsible citizens"* (quote referring to 'interacting in heterogeneous groups'; teacher P9)

At upper secondary school level, two lines of argumentation were found:

- Decision related to the relevance of the competence in the students' further education: *"Important for life at university"*; *"important for the Matura thesis"* (both quotes referring to 'acting autonomously' teachers S7; S11)
- Decision related to the general relevance of the competence: *"Important for life"* (referring to 'interacting in heterogeneous groups'; teacher S2).

The units trialled were characterized by three criteria: The dimension(s) of openness; the inquiry activities enacted in the units; and the competences assessed. Openness refers to the idea that in inquiry-based education, not all aspects are pre-defined but some decisions are left to the students. These decisions may concern different dimensions. The dimensions of openness were conceptualized after Priemer (2011; see sub-chapter 3.1 for details). The analysis of the trials in this study shows that all dimensions of openness were covered by at least a few trials at both school levels. Whereas only few inquiry units were open in terms of 'content' and 'strategies', more inquiry units were open in terms of 'methods'. Almost all inquiries were open in terms of 'solution' and 'solution process'. The differences between school levels were small. Many trials covered more than one dimension of openness: The peak at primary school was around units which were open in three dimensions. The distribution at upper secondary school had a maximum at 'open in two dimensions' and another maximum at 'open in five dimensions'.

Looking at the inquiry activities enacted in the units, it appears that all activities as defined in Bell et al. (2010; see sub-chapter 3.1 for details) were part of at least one trial at both school levels. However, huge differences in the frequency occur: Whereas 'orienting and asking questions'; 'hypothesis generation'; 'model'; and 'prediction' were rarely part of the inquiries at both school levels, 'planning'; 'investigation'; 'analysis and interpretation'; and 'communication' were frequently enacted at both school levels. 'Conclusion and evaluation' was often part of the inquiries at upper secondary school but not at primary school. Looking at the number of inquiry activities per unit, the peak of the primary school units is around four inquiry activities. At upper secondary school, there is no clear peak, but most trials included between 4 and 6 activities.

Both domain-specific and transversal competences are ascribed to inquiry-based education (see sub-chapter 3.1 for details). In this study, the conceptualisation of domain-specific competences that are fostered by inquiry from Bell et al. (2010) were taken as a basis. The results show that all domain-specific competences were assessed at least once in the trials. However, there are differences in the frequency of occurrence: At primary school, 'orienting and asking questions', 'model', 'conclusion and evaluation', as well as 'prediction' was not assessed at all. 'Hypothesis generation', 'planning', and 'analysis and interpretation' were rarely assessed. By far the most-assessed competences in the primary school trials were 'investigation' and 'communication'. At upper secondary school, 'hypothesis generation' was not assessed but all other competences were. 'Orienting and asking questions', 'model', 'conclusion and evaluation', and 'prediction' were rarely assessed. 'Planning', 'investigation', and 'analysis and interpretation' appeared at a moderate frequency. By far the most-assessed competence was 'communication'. The results also show that trials with one domain-specific competence assessed were most frequent at both school levels. Two, three or four competences occurred less frequently. At upper secondary school, there was a small number of trials with 5 or 6 competences assessed.

When deciding about what domain-specific competences to assess, the decision-making process of the teachers seems to take place on different levels: At primary school level, the most frequently mentioned line of argumentation included no explicit decision. Instead, the teachers explained that the competences for assessment emerged naturally during the preparation of the unit. Less frequently, the teachers brought up resource-based decisions. Finally, some teachers chose a particular competence because they thought it was important for science education. At upper secondary school, the two most commonly mentioned lines of argumentations were that a particular competence was considered important for the students' further career or generally important in science education. Less frequently, the decision was taken based on the students' abilities.

The analysis of the transversal competences assessed was based on the conceptualizations from OECD (2005b, see sub-chapter 3.1 for details). The results show that transversal competences were assessed in most trials at primary school and in half of the trials at upper secondary school. They were always assessed in combination with at least one domain-specific competence. The teachers' reasons for deciding on a particular transversal competence were driven by the perceived relevance of this competence at both school levels.

7.2.2 Description of the formative assessment activities trialled

In this section, the formative assessment methods trialled will be characterized. The 34 cases that matched the criteria from sub-chapter 5.4 will be included in this part of the results.

Communicating the criteria

At some point of the trials, the students had to be introduced to the criteria of assessment (see sub-chapter 3.2 for theoretical background). In most of the trials (11 out of 14 trials at primary school; 16 out of 20 trials at upper secondary school; see Figure 19), the criteria of assessment were handed out and introduced at the beginning of the unit. Some teachers formulated the criteria as questions; others used rubrics or lists of criteria. In two cases at primary school, the assessment criteria were not pre-defined by the teacher but elaborated together with the students during the unit. In a few cases (one trial at primary school, four trials at upper secondary school), the assessment criteria were not explicitly introduced but were clear from the context. One of these examples was the unit on the construction of pendulum clocks at primary school (see description of cases in appendix A7; teacher P6): The teacher expected that it was clear for the student that the focus of the reflective discussions would be on the construction process.

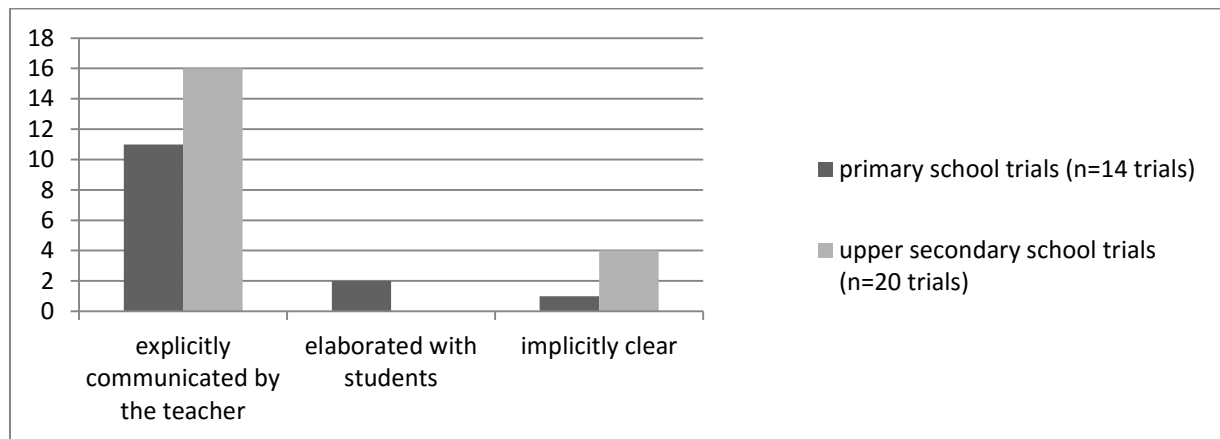


Figure 19: Introduction of assessment criteria.

Data sources for diagnosis

The formative assessments in the trials were based on a number of sources of data. These sources of data were grouped into four categories (see sub-chapter 3.2): Written student data (such as lab journal entries, reports, and similar); artefacts and models; oral student data (such as student conversations, presentations etc.); and observational data (the assessor observing a student's or several students' behaviour). In the trials in the study, all four sources of data were used. However, the frequency of use differed between methods and also between school levels (see Figure 20). Written student data and observational data were generally more common than models/ artefacts and oral student data. Written student data was most common at the upper secondary school trials, observational data was most common at the primary school level trials.

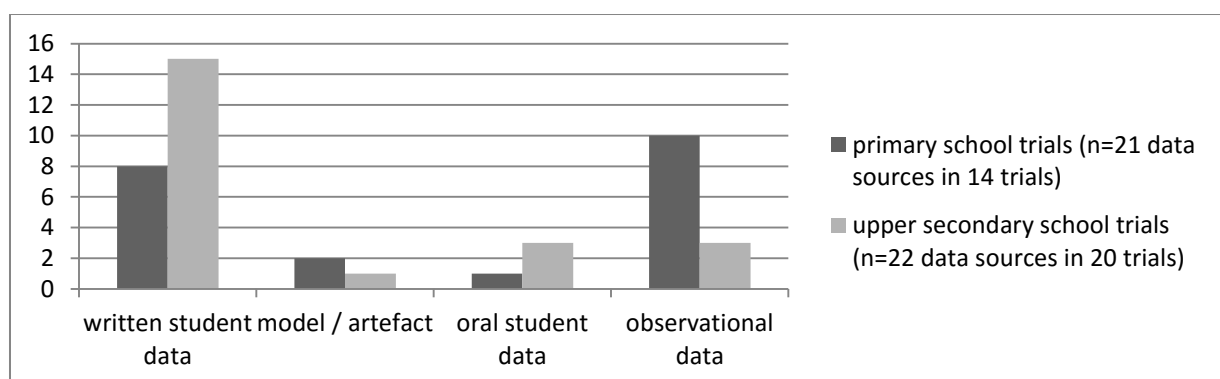


Figure 20: Data sources for diagnosis.

Investigating the sources of data in the individual trials in more detail, the number of data sources as a basis for assessment per trial was analysed (see Figure 21). The analysis showed that in all trials, either one or two data sources were used. In the primary school trials, the two possibilities occurred equally frequent whereas in the upper secondary school trials, the use of one data source was more frequent. An example of using two types of data on student learning in the same formative assessment activity was the first trial of primary school teacher P8 (see appendix A7 for details): The respective inquiry focussed on the movement of different animals. The teacher based her written assessment on both the observation of the students' behaviour (basis to assess the interaction in groups; acting autonomously) and a draft report of the inquiry (basis to assess communication of results of the inquiry).

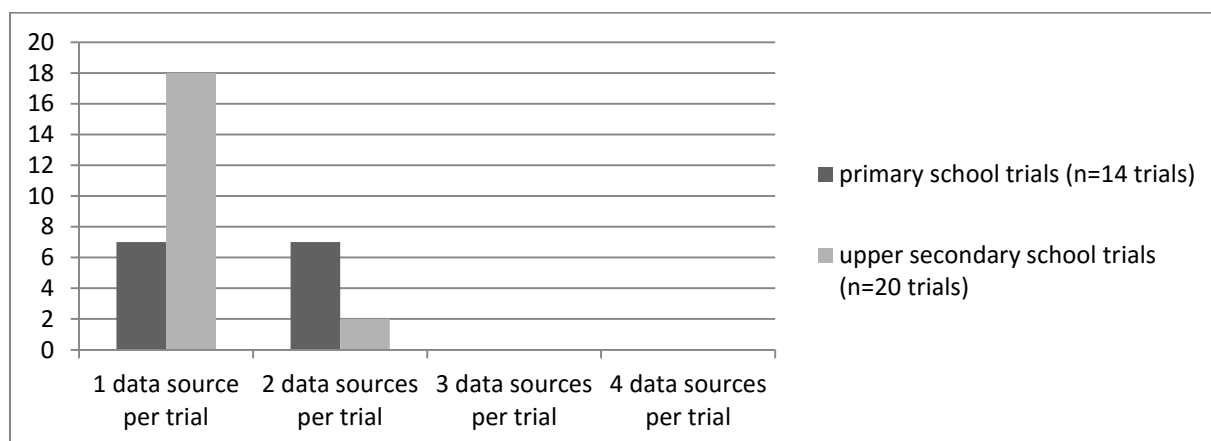


Figure 21: Number of data sources per trial at primary and at upper secondary school level.

Assessment methods

The following methods of formative assessment were trialled in the study: Written teacher assessment, peer-assessment and self-assessment (see sub-chapter 3.4 for theoretical background). The number of trials per method is displayed in Figure 22.

Even though the teachers were supposed to trial one assessment method per semester, some teachers decided to embed two or more methods in the same unit which resulted in more assessment methods than trials as displayed in Figure 22. In more detail, four trials at primary school involved more than one assessment method: In three trials, teacher- and peer-assessment were combined whereas in one case, peer-assessment was combined with self-assessment (this case is described in more detail in the illustrative examples; sub-chapter 6.3). At upper secondary school, two trials had more than one assessment method embedded: One trial involved teacher- and self-assessment whereas the other trial consisted of teacher- and peer-assessment. In all trials where more than one assessment methods were embedded, these methods were used to assess different competences. One of the primary school teachers (P9) explained this effect as follows: *“The science units are particularly suitable to assess students formatively. So I have to find ways to get grades for the annual reports but also to support students individually. I am continuously searching for and trying out different approaches for the supportive part. This is probably why I have rather many formative assessment activities in the same unit: I usually start with the topic, with the content, but as soon as I recognise situations suitable for formative assessment in the planning, I embed corresponding activities.”*

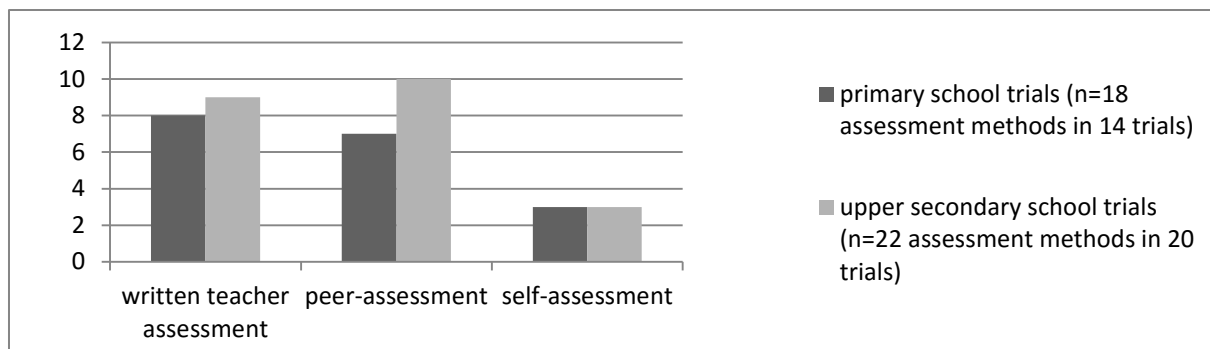


Figure 22: Assessment methods.

The methods were put into practice in different ways, as will be introduced in the subsequent paragraphs: The written teacher assessment was provided either in the form of filled-out rubrics, in the form of open comments, or in a combination of both. In some cases, the teachers added formative comments to graded reports. In the illustrative example on the growth of chicks in sub-chapter 6.1, the teacher worked with open comments: She had two pre-defined criteria and provided written advice as appropriate for the individual students.

The peer-assessment was also trialled in different varieties: Some teachers had the students produce artefacts and providing feedback in groups, others implemented individual feedback. In a number of cases, the two types were mixed (individual feedback on group artefact or vice-versa). In one of the first meetings, peer-assessment in the form of whole-class discussions (meaning that all artefacts are laid out in the room and all students assess all artefacts) was brought up by a primary school teacher. The idea was later adopted by a number of teachers from both school levels. All teachers embedded the peer-assessment reciprocally, that means that all students of a class acted both as assesses and assessors. All teachers provided scaffolds in the form of criteria for the peer-assessment (see paragraph above). In the illustrative example on the revision of lab-reports in sub-chapter 6.2, the criteria of assessment were not provided in a plain list but formulated as questions in order to guide the students' focus of assessment.

The self-assessment was put into practice in the following ways: Reflection sheets filled out individually (1 case at primary school which is described in the illustrative example on the soil profiles in sub-chapter 6.3; 1 case at upper secondary school); reflective discussions in groups of students (2 cases at primary, 1 case at upper secondary school); reflective discussions between the teacher and one student (1 case at upper secondary school). One of the reflective discussions in groups was the first trial of primary school teacher P6 (see appendix A7 for details): She let her fourth-grade students construct a pendulum clock. The students were given the construction kit but no instructions on how to combine the toothed wheels, hands etc. For this task, the students worked in groups of three children. The reflective discussions focussed on the construction process. Every group of students sent a delegate to the discussion round. The other members of the group listened to the reflective discussions. The first discussion was initiated by the teacher who asked: "what is the pendulum good for, how can it be connected to the rest of the clock?" The students expressed their ideas and came up with hints for the next steps in the construction process. In the subsequent discussions, the students brought up questions that were relevant at that moment.

Figure 23 reveals what kind of competences (domain-specific vs. transversal) was assessed with which assessment method. There are no clear differences between the two school levels. Instead, the two combinations written teacher assessment for assessing domain-specific competences and peer-assessment for assessing domain-specific competences appear to be the most frequent in the study. As mentioned earlier, transversal competences were assessed less frequently in the study. There is no clear trend by which assessment method they would be assessed most frequently.

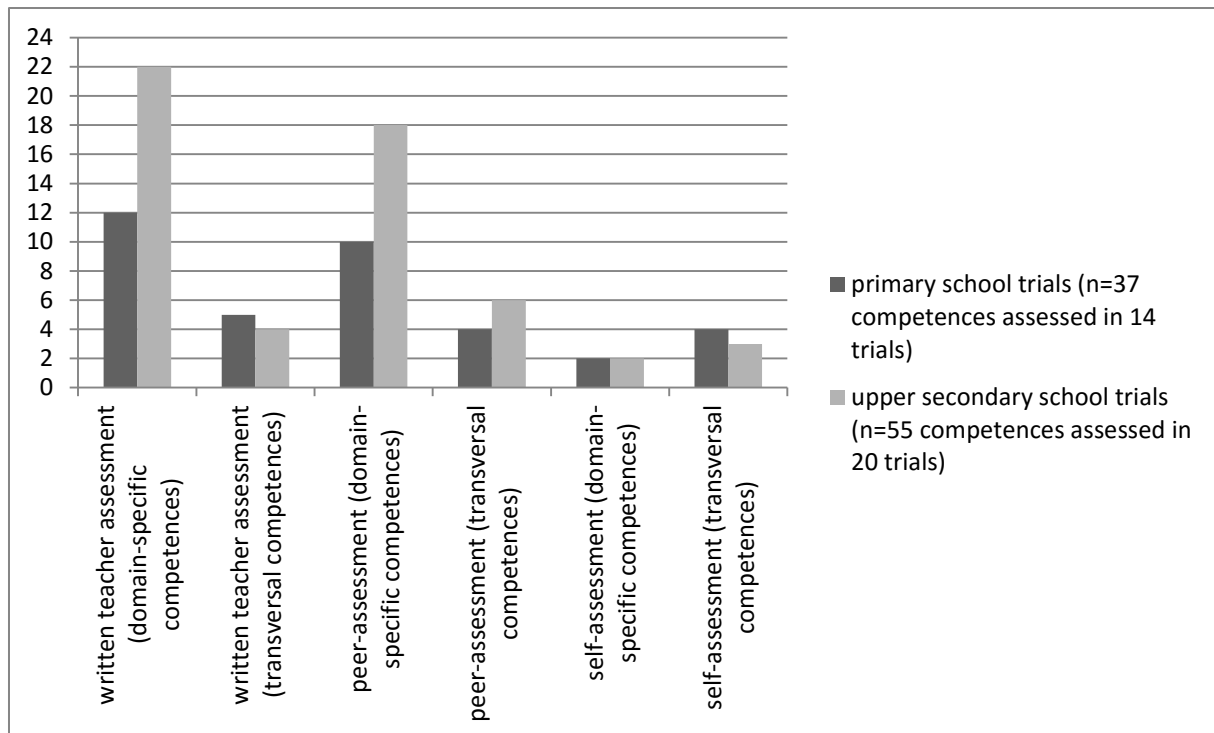


Figure 23: Domain-specific and transversal competences assessed by the different assessment methods.

In the individual interviews, the teachers were asked about the reasons for the choice of a particular assessment method. At primary school level, the respective answers were grouped in the following three lines of argumentation:

- No explicit decision but “*organic growth*” (teacher P9); «*appeared suitable*» (teachers P5; P6)
- Decision related to the teachers’ self-efficacy: «*Thought I would be able to manage it*» (teacher P4)
- Decision related to the students’ motivation: «*Thought the students would like it*» (referring to peer-assessment; teacher P2)

At upper secondary school level, the following four lines of argumentation emerged from the teachers’ answers:

- No explicit decision but “*just emerged*” (teacher S4); «*it appeared suitable*» (teacher S11)
- Decision related to the teachers’ self-efficacy: «*Appeared doable and convenient*» (teacher S7); «*used the method in the round before*» (teacher S10)
- Decision related to the development of generic competences: «*The method will improve the students’ reflective skills*» (referring to self-assessment; teacher S2)
- Decision related to organisational issues: «*Because of the size of the class*» (referring to peer-assessment; teacher S1); «*will help the students to be right back in the topic after their exchange program*» (referring to written teacher feedback; teacher S5).

Means of engaging with the feedback

The teachers in the study were asked to embed their formative assessment methods in a way that there was an opportunity for the students to engage with the feedback received. The teachers found two ways of ensuring this: Either by giving the students the possibility to revise their original artefacts based on the feedback or by setting up a similar task or situation (for example a subsequent lab report at upper secondary school) to which the feedback could be transferred. In the illustrative examples in chapter 6, both varieties occur: In the example with the chick journal (sub-chapter 6.1), the students were expected to transfer the feedback received to the subsequent journal entry rather than to revise the initial entry. In the example on stationary waves (sub-chapter 6.2), on the other hand, the students had the opportunity to revise their draft lab reports based on the peer-assessment received before the teacher graded the revised reports.

The frequency of occurrence of the two possibilities can be found in Figure 24. The results show that in the primary school trials, the engagement with the feedback was likely to be a revision of the original artefact whereas at upper secondary school, a transfer to a similar subsequent task or situation was more common.

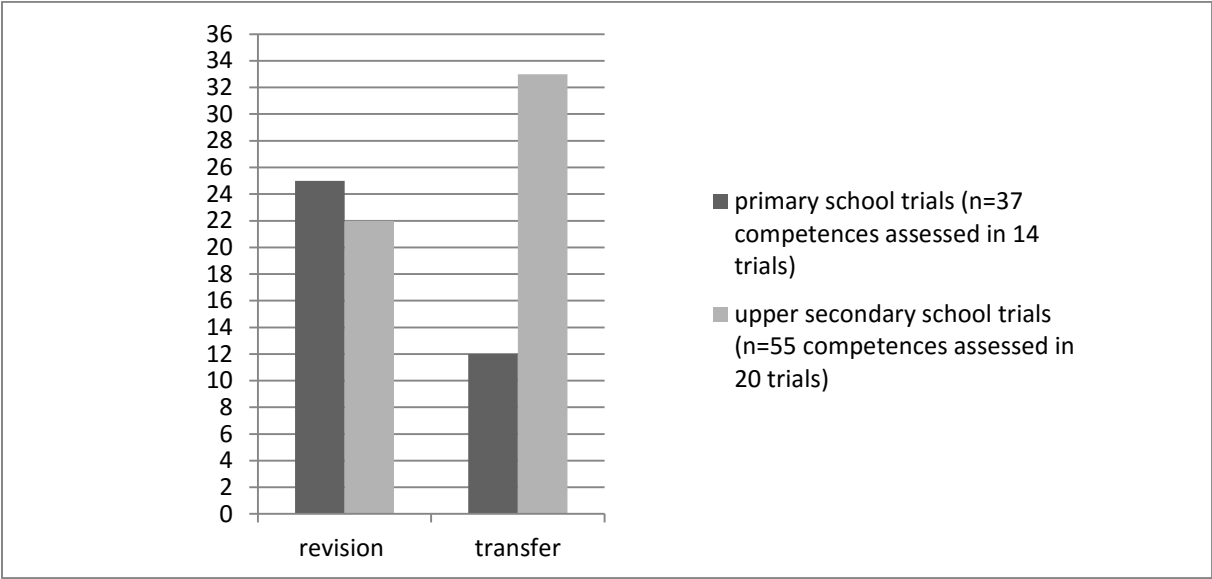


Figure 24: Means of engaging with the feedback in the primary- and in the upper secondary school trials.

As Figure 25 shows, the means of engagement with the feedback (revision or transfer) does not only seem to relate to the school level but also to the competence assessed: With domain-specific competences, revisions of the original artefacts were more common than with transversal competences; particularly at primary school level. With transversal competences, the transfer to subsequent activities and situations was more frequent.

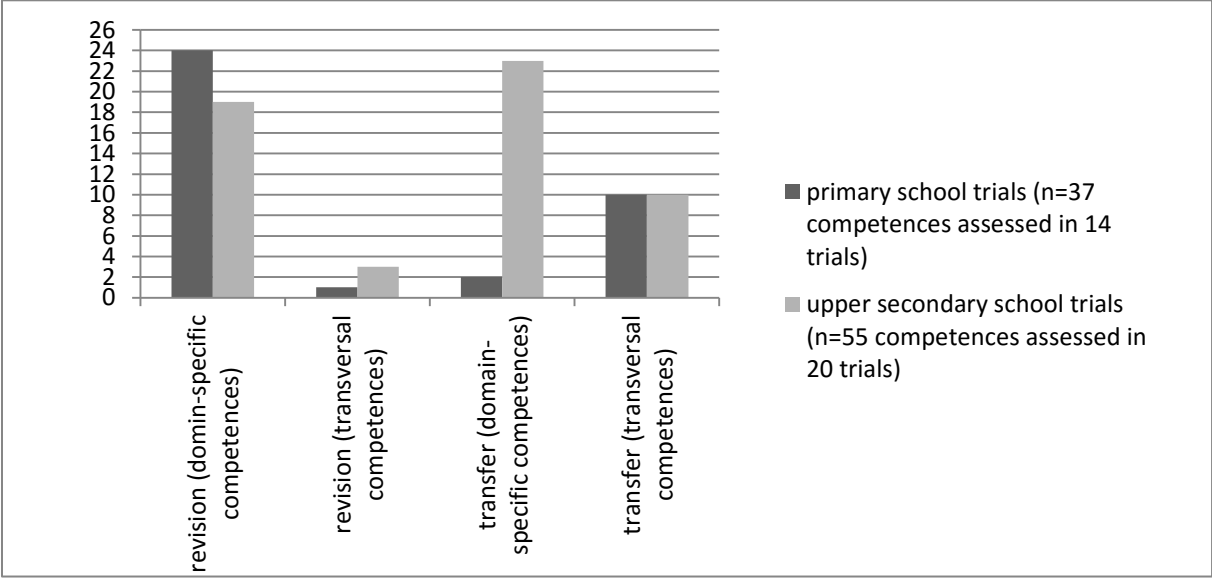


Figure 25: Means of engaging with the feedback differentiated by competence.

Cycles

As laid out in sub-chapter 3.2, formative assessment can be classified in terms of the length of a cycle. This length describes the timespan within which evidence on student learning is collected and the possibility to adapt this learning (e.g. after the feedback has been received). In the illustrative examples in chapter 6, all three varieties as defined in sub-chapter 3.2 occur: In the example with the chick journal (sub-chapter 6.1), the students had the opportunity to use the feedback received from their teacher the next day, resulting in a short cycle. In the example on stationary waves (sub-chapter 6.2), on the other hand, the students had the opportunity to revise their draft lab reports a week after receiving the peer-assessment, resulting in a medium cycle. In the example with the soil profile in sub-chapter 6.3, the feedback should be transferred to the next half-day excursion to the local forest which took place next month, resulting in a long cycle.

Figure 26 displays the cycle lengths of the trials at primary and upper secondary school level. In total, 34 trials were classified (14 cases from primary school, 20 cases from upper secondary school). In a number of trials where different competences were assessed, the cycle length differed between these competences. Therefore, the analysis was conducted competence-based rather than trial-based.

Apparently, the primary school teachers generally chose to implement their formative assessment activities in a shorter time span than the teachers at upper secondary schools. None of the trials from primary school was classified as long (4 weeks or longer) whereas this applied for 10 of the upper secondary school cases.

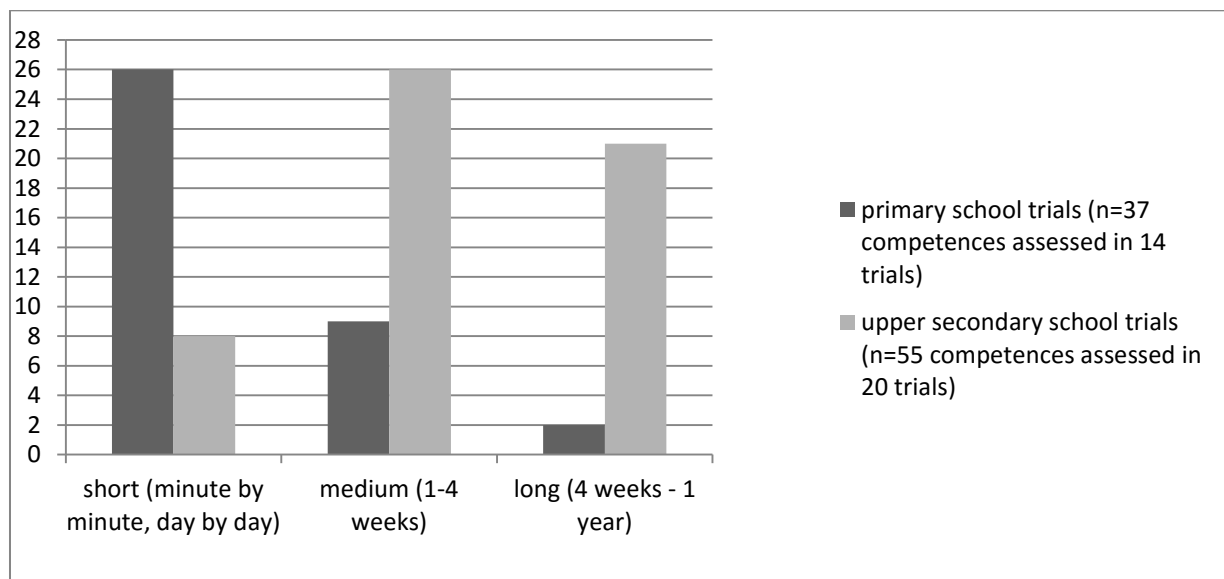


Figure 26: Cycle length of the trials.

Figure 27 reveals that in the trials of the study, there was not only a relation between cycle length and school level but also between cycle length and competence assessed: The domain-specific competences were assessed in the context of shorter cycles than the transversal competences at both school levels (e.g. the teachers in the study intended the feedback on a students' transversal competences to be used over a longer time span than the feedback on domain-specific competences).

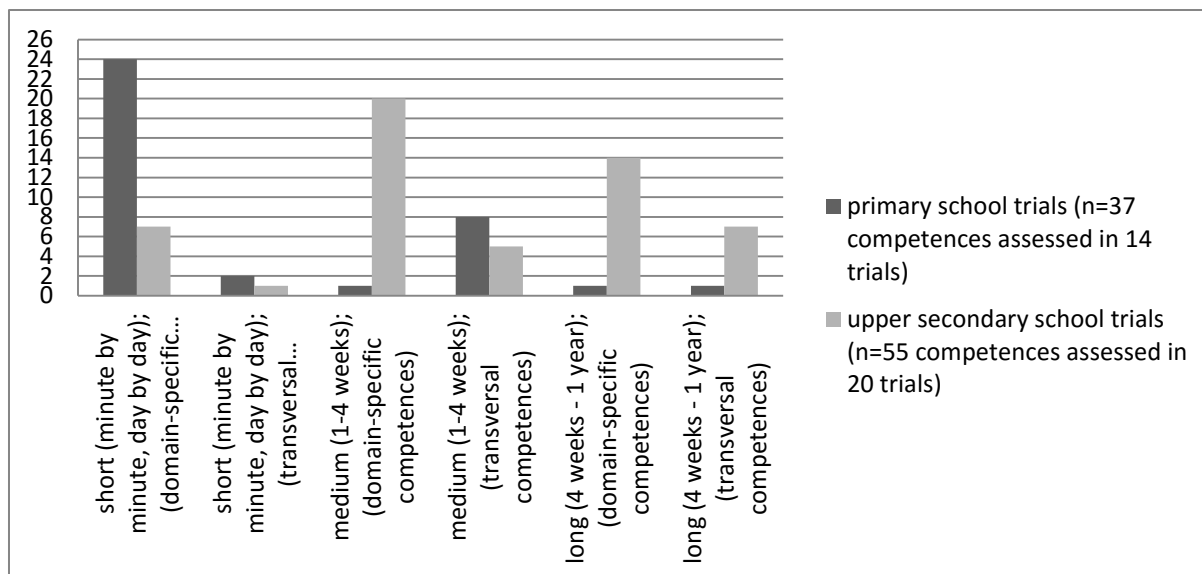


Figure 27: Cycle length of the trials differentiated by competences.

The teachers' formative assessment activities were characterized in terms of different aspects: The communication of the criteria, the data sources used for diagnosis, the assessment methods, the means of engaging with the feedback, and the length of the assessment cycles.

As part of the formative assessment, the assessment criteria were introduced in the trials. Three ways of introduction were found in the study: Most frequently at both school levels, the assessment criteria were pre-defined and explicitly communicated by the teacher. At primary school level, the assessment criteria were elaborated with the students in two cases. In a few cases at both school levels, the criteria were implicitly clear (for example in some of the cases where documenting experimental results in the lab journal are regular part of the lab lessons).

For diagnosis, a number of types of data on student learning were used. Amongst those, observational data was most frequent in the primary school trials whereas written student data was most common in the upper secondary school trials. In the primary school trials, one or two sources of data were used whereas in the upper secondary school trials, the use of only one source of data was most common.

Three formal formative assessment methods were trialled (written teacher assessment; peer-assessment; self-assessment) several times at both school levels. At both school levels but more often at primary school level, more than one assessment method was embedded in one trial even though that was not part of the teachers' task. The relation between assessment methods and competences assessed was analyzed, but no clear pattern emerged: All three assessment methods were used to assess both domain-specific and transversal competences.

In the individual interviews, the teachers were asked how they chose a particular assessment method for their trials. A number of decision-making processes could be revealed: At both school levels, some teachers answered that there was no particular reason for their choice but that the formative assessment just appeared suitable in the context of a particular situation. A second line of argumentation appearing at both school levels was related to the teachers' confidence to work with a particular method. At primary school level, the third reason was the students' motivation to work with a particular method. At upper secondary school level, the learning benefits of using the method itself (reflective skills that develop from self-assessment) and organizational issues were also mentioned.

As a next step, the means of engaging with the feedback received was analyzed. The two options mentioned in the literature (see sub-chapter 3.2) are revision of the original artefact/activity or transfer to a subsequent artefact/activity. The results from the study showed that at primary school, the feedback was more likely to be used for revision whereas at upper secondary school, the feedback rather enhanced the subsequent work (e.g. the next lab report). A second effect which was visible was that feedback on domain-specific competences was more often used for revision whereas feedback on transversal competences was more likely to be transferred to subsequent units.

The cycle length of the formative assessment activities trialled (see sub-chapter 3.2) was analyzed. The results showed that in the study, the primary school teachers' cycles were typically shorter (minute by minute, day by day) whereas the upper secondary school teachers had longer cycles (1-4 weeks or 4 weeks to 1 year). A second effect that was visible in the study was that feedback on domain-specific competences was typically used in shorter cycles whereas feedback on transversal competences was more likely to be used in longer cycles.

7.2.3 Problems in the trials

In this section, the 19 cases that did not match the criteria from sub-chapter 5.4 will be characterized. The section will be structured along the four criteria from sub-chapter 5.4: Conduction of trial; sufficient documentation of trial; inquiry nature; formative nature of assessment (see Figure 28).

Three teachers from primary school and four teachers from upper secondary school did not implement any formative assessment method in their inquiry teaching in one of the semesters. All of them offered time issues as reasons.

Three teachers from primary school did implement an assessment method in their inquiry teaching and told about it in the group discussions at the end of the semester. But they did not hand in lessons plans and teaching materials. These cases were classified as non-matching the criteria from sub-chapter 5.4 because the documentation from the group discussions solely was insufficient for analysis.

One teacher at upper secondary school (S8) and a teacher at primary school level (P9) did implement a formative assessment method but not in the context of inquiry-based education as defined in sub-chapter 5.4. The teacher from upper secondary school was aware of the fact she was not following the instructions given from the project. She mentioned that she did not have time to do inquiry and formative assessment so she tried her best to at least do whatever seemed possible.

Three teachers from primary school and four teachers from upper secondary school implemented no formative assessment method. One of the primary school teachers (P1) did not communicate the criteria of assessment before the diagnosis and feedback processes started. The second primary school teacher (P4) graded her students according to assessment criteria but did not provide them with any opportunity to make use of this information. The third primary school teacher (P6) focussed on issues related to the documentation of the students' learning progress in all three rounds of implementation. In one trial, she decided to make *"the learning of the students visible with mind maps"*. However, this idea was easier to realize with content knowledge than with inquiry competences which is why that trial did not match the criteria from sub-chapter 5.4. At upper secondary school, one teacher (S11) concentrated on the inquiry part of the unit and simply forgot about the assessment. He became only aware of this in the group discussion at the end of the semester when the other teachers asked about it. Another teacher (S8) did not communicate the criteria of assessment before the diagnosis and feedback process started. The two remaining teachers at upper secondary school did not provide the students with an opportunity to use the feedback they had received.

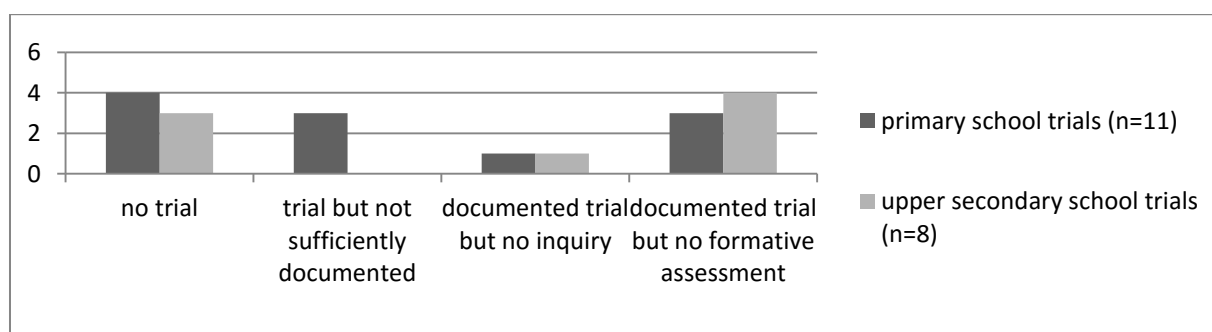


Figure 28: Problems in the trials

Nineteen out of 53 cases collected in the study did not match the criteria as specified in chapter 5.4. These criteria were: (1) Conduction of trial; (2) sufficient documentation of trial for analysis; (3) trial in the context of inquiry-based education; (4) trial involving formative assessment.

Seven teachers from both school levels did not manage to trial anything even though they were paid for it and motivated to do it. They all said that they did not find time during the semester to conduct a formal

formative assessment activity in the context of inquiry due to the extensive curriculum and due to the many other activities (such as teachers' military service or student exchange programs).

Three primary school teachers' trials were not documented in a way they could be analyzed which is a methods issue.

In the case of a teacher from primary school and a teacher from upper secondary school, the problem was not with formative assessment but with inquiry. The teacher from upper secondary school was aware of the fact that she was not following the instructions given for the study whereas the teacher at primary school was not.

A few teachers from both school levels did not trial formative assessment activities. One of the teachers simply forgot the formative assessment part because he was so busy with the inquiry. The other teachers either did not clarify the assessment criteria at the beginning of the formative assessment activity or they did not provide the students with the opportunity to make use of the feedback they received. Some of them were not aware of the issues with their trials whereas others mentioned them during the group discussions.

7.3 Results on research question 3: Teachers' and students' evaluations of the methods trialled

In the third sub-chapter, the data relating to research question 3: *How do the teachers and the students evaluate the formative assessment methods trialled?* will be presented. So the aim of this sub-chapter is not to know what the teachers did (as in the second sub-chapter) but to reconstruct what both the teachers and the students thought about the methods trialled.

In the beginning of the sub-chapter, the teachers' perceptions of benefits and challenges related to the different assessment methods which they trialled will be presented (sections 7.3.1 and 7.3.2). This is investigated by analysing the teacher evaluation forms as well as individual teacher interview data. In the later part of sub-chapter 7.3, the evaluation form for students will be analysed (sections 7.3.4 and 7.3.5). Both the teachers and the students were asked to suggest possible means of support that might facilitate the formative assessment practices. The respective answers will be enclosed in sections 7.3.3 for the teachers and 7.3.6 for the students.

7.3.1 Usability of methods for certain school levels from the teacher perspective

In the teacher evaluation form (see section 5.2.4), all teachers were asked to what extent they considered the assessment method they trialled usable for assessing competences in the context of inquiry-based education at their school level. The two tables below provide an overview of the answers given from the teachers who successfully trialled the respective methods (see sub-chapter 5.4). As pointed out in section 7.2.2, some of the teachers trialled more than one assessment method in the same trial. Therefore, the number of trialled methods differs from the number of trials.

Table 24 displays the results of the primary school teachers. The teachers agreed that all the assessment methods they trialled are 'rather useful' or 'very useful' for their school level. Potential reasons for these evaluations can be found in section 7.3.2.

Table 24: Primary school teachers (in n=18 methods trialled in 14 trials): Usability of methods. 4=very useful; 1=very useless.

Usability of assessment methods	4	3	2	1
Written teacher assessment	6	2		
Peer-assessment	2	4	1	
Self-assessment	1	2		

Table 25 displays the results of the upper secondary school teachers. The teachers agreed that all the assessment methods they trialled are 'rather useful' or 'very useful' for their school level, and potential reasons for these evaluations can be found in section 7.3.2.

Table 25: Upper secondary school teachers (in n=22 methods trialled in 20 trials): Usability of methods. 4=very useful; 1=very useless.

Usability of assessment methods	4	3	2	1
Written teacher assessment	6	3		
Peer-assessment	4	6		
Self- assessment	1	2		

All three formative assessment methods trialled in the study were considered usable at their respective school levels by the teachers in the study. Reasons for these evaluations can be found in the subsequent sections.

7.3.2 Benefits and challenges of different methods of formative assessment as mentioned by the teachers

In the evaluation forms of the successful trials (see sub-chapter 5.4; n=34 evaluation forms from 20 teachers) but also in the individual interviews (n=16 interviews with 14 teachers), the teachers were asked what benefits and challenges they perceived related to the specific assessment methods in the context of inquiry-based science education at their school levels. The themes that emerged when the teachers spoke and wrote about advantages and disadvantages of the different assessment methods will be introduced first, since the themes were the same for both positive and negative evaluations across school levels. Afterwards, more specific results on each of the formative assessment methods will follow.

Themes covered by the teachers speaking about benefits and challenges of the formative assessment methods trialled

From the teachers' evaluations of the different assessment methods, ten themes emerged: Embedding formal formative assessment methods in inquiry-based science education; diagnosis of students' levels of achievement; content of the feedback; role of the teacher; documentation; use of the feedback by the students; learning effects; social and motivational effects; relation between formative and summative assessment; and effort needed. Each of the themes was defined as a category with a number of sub-categories subsumed. Some of the categories were used to code either benefits or challenges of the different assessment methods, others were used for both benefits and challenges. The resulting coding system will be displayed in Table 26.

Table 26: Coding system for benefits and challenges of assessment methods trialled as perceived by the teachers.

Category	Sub-categories
embedding formal formative assessment methods in inquiry-based science education	<ul style="list-style-type: none"> - long-term planning aspects - short-term planning aspects
diagnosis of students' levels of achievement	<ul style="list-style-type: none"> - time pressure - pre-defined criteria - quality of diagnosis - individuality
content of feedback	<ul style="list-style-type: none"> - timing - focus - quality in terms of content - quality in terms of language and vocabulary - relation between assessor and assessee - potential for enhancement of learning
role of the teacher	<ul style="list-style-type: none"> - responsibility for student learning - workload for teacher - capacity for individual support
use of the feedback	<ul style="list-style-type: none"> - eagerness of recipients - understanding of feedback - transfer
learning effects	<ul style="list-style-type: none"> - scientific concepts - science-specific competences - transversal competences - self-regulated learning
social and motivational effects	<ul style="list-style-type: none"> - relation between teacher and student - classroom climate - motivation
documentation	<ul style="list-style-type: none"> - record of feedback for students - record of feedback for teacher - communication with parents
relation between formative and summative assessment	<ul style="list-style-type: none"> - relevance of formative assessment - check-like character
effort needed	<ul style="list-style-type: none"> - time - practice

Embedding formal formative assessment methods in inquiry-based science education

This category covers benefits and challenges related to when and how to integrate formal formative assessment methods in the semester- or lesson plan. Two sub-categories emerged from the teachers' quotes: Long-term planning aspects and short-term planning aspects. Both sub-categories contained quotes that point towards organisational aspects and quotes that point towards educational aspects.

Diagnosis of students' levels of achievement

This category covers benefits and challenges related to the diagnosis of students' levels of achievement. Four sub-categories emerged from the teachers' quotes: Time pressure for the teacher within the lesson; pre-defined criteria; quality of diagnosis; and individuality.

Time pressure for the teacher within the lesson refers to the time available to diagnose the individual student's level of achievement. One of the teachers (S7), for example, put it like this: *"One of the advantages of the written teacher assessment is that I <the teacher> can take my time to decide about the exact content of my comments."*

The second sub-category, the pre-defined criteria, refers to advantages and challenges that emerge from the fact that the students' levels of achievement are assessed along a set of criteria that have been decided upon beforehand. Respective examples of aspects mentioned by the teachers include the objectivity of the assessment and the possibility to work on the same set of criteria with a team-teaching partner. Teacher P5, for example, said: *"One of the advantages of the pre-set criteria is that my team-teaching partner can also focus on the same criteria in her lessons. Or if she replaces me in one of my lessons, she immediately knows what to concentrate her support on."*

The third sub-category, the quality of the diagnosis, reflect the teachers' evaluations of whether a particular method of formative assessment led to an accurate and correct diagnosis of the students' levels' of achievement. Teacher P1 addressed this aspect when speaking about peer-assessment: *"Some students perceive their peers' level of achievement different than I would. This leads to a different diagnosis."*

The last sub-category, the individuality in the diagnosis, subsumes teacher quotes on the student-specific nature of the diagnosis. Examples include the evaluation to what extent formative assessment makes the individual learning process visible or to what extent the diagnosis is student-specific rather than a general impression on the level of the class as a whole.

Content of feedback

This category covers quotes on the content of the feedback in the context of the formative assessment methods. Six sub-categories emerged from the teachers' quotes: Timing; focus; quality in terms of content; quality in terms of language and vocabulary; relation between assessor and assessee; potential for enhancement of learning.

In the first sub-category, quotes on the timing of the feedback are subsumed. These basically cover the question whether a particular method provides immediate feedback or delayed feedback (e.g. the week after). An example was a teacher's evaluation (P4) who considered it an advantage that the peer-assessment comes immediately and not *"only after I [teacher] have looked through all the worksheets a week or two later"*.

The fourth sub-category summarizes quotes on the quality in terms of language and vocabulary used in the feedback in different types of formative assessment. An example is the impression that peer-assessment is easy to understand and apply because for students the language and vocabulary used is familiar to them. In this context, it was mentioned (teacher S10) that *"peer-assessment is different from teacher assessment. It is much more recipe-like: do this, do not do that, and so on. Assessment from the teacher contains much more explanations."*

In the fifth sub-category, the relation between assessor and assessee was focussed on. Apparently, this differs between the different assessment methods and may result in a difference in the acceptance of the feedback. The point was made that criticism from peers is easier to accept and taken more serious than criticism from the teacher. Peer-assessment was therefore considered particularly effective to plan the students' further learning by some of the teachers. It was also noted that the inhibition level to ask back to the assessor in case the feedback was not understandable was low in a peer-assessment setting.

The last sub-category has quotes on the potential of a particular assessment method for the enhancement of student learning subsumed. Different mechanisms of support of the student learning, depending on the assessment methods, were identified by the teachers: The impression that feedback raises the students' awareness of the assessment criteria, or that it allows the students to easily draw conclusions for their further learning.

The role of the teacher

This category covers quotes on the tasks and responsibilities of the teacher in the context of self- and peer-assessment where, naturally, the students were active in the diagnosis and the feedback parts of the formative assessment. Three sub-categories emerged from the teachers' quotes: The responsibility for student learning; the workload for the teacher; and the teacher's capacity for individual support.

In the first sub-category, the issue of who is responsible for the students' learning in the context of self- and peer-assessment is covered. Teacher S3, for example, put it like this: *"Peer-assessment signals to the students that they are responsible for their learning, not me. It is a way to show them that they become autonomous, independent learners."*

The second sub-category subsumes quotes on the workload for the teacher under the circumstances of self- and peer-assessment. Teachers from both school levels mentioned that peer-assessment reduces the workload of them.

Related to the last point is the third sub-category: The teacher's capacity for individual student support. Some teachers mentioned that self- and peer-feedback provides the opportunity for the teacher to focus on problems of individual students while *"the whole lot is busy assessing each other without me"* (teacher P4).

Learning effects

This category covers quotes on the effects of the different formative assessment methods on the students' levels of performance. Four sub-categories emerged from the teachers' quotes: Scientific concepts; science-specific competences; transversal competences; and self-regulated learning.

The sub-category 'scientific concepts' encloses quotes referring to the students' improving on their knowledge in science. This could include the idea that they engaged with the content once more or that they revised a particular concept in the course of formative assessment activities.

'Science-specific competences' as the second sub-category was used to code references to the improvement of the students in terms of science-specific competences such as developing hypotheses, collecting measurement data, and documenting investigations.

Similarly to the afore-mentioned sub-category, 'transversal competences' was used to code references to the students' improvement in terms of transversal competences, such as providing feedback or collaborating with peer students.

The sub-category 'self-regulated learning', finally, encloses quotes referring to the students' improvement in the self-regulation of their learning.

Social and motivational effects

This category focusses on social and motivational effects of the different assessment methods as perceived by the teachers. Three sub-categories were formed: Effects on the relation between teacher and student; effects on the classroom climate (among students); and effects on the motivation of the individual student.

The first sub-category, 'effects on the relation between teacher and student' encompasses the teacher quotes stating that due to the formative assessment activities, the relation between themselves and the individual students changed. These changes could be positive, meaning that the relation between the student and the teacher improved; or negative, meaning that the relation between the student and the teacher worsened.

'Effects on the classroom climate', the second sub-category, was used to code teacher quotes saying that formative assessment changed the relation between the students in the classroom. This change could be to

the better or to the worse. Teacher S10 said: *“Assessing their peers and this way providing advice and expressing their appreciation to their work is a way for students to show their respect to each other. In that sense, peer-assessment can change the classroom climate.”*

Finally, ‘effects on the motivation of the individual student’, the third sub-category, encloses teacher quotes expressing that the motivation of the individual student changed due to a formative assessment activity. Again, this change could be positive, meaning that the student motivation improved; or negative, meaning that the student motivation worsened. Teacher P8 put it like this: *“My feedback shows the individual student that I noticed him or her and his or her work. It motivates them to proceed.”*

Documentation

This category covers quotes on the documentation of the feedback. Three sub-categories emerged from the teachers’ quotes: The record of the feedback for the student; the record of the feedback for the teacher; and the documentation of the feedback as a means of communication with the parents.

The first sub-category subsumes the teachers’ impressions on whether the students have a record of the feedback received in the context of a particular assessment method. One teacher (S1) said: *“Written teacher assessment is good simply because it cannot be forgotten easily like oral feedback.”*

The second sub-category focusses on the question whether the teachers have a record of the feedback provided to the students in the context of a particular assessment method.

The last sub-category covers thoughts on the usability of the assessment methods to provide the parents with an impression on their childrens’ learning. One of the primary school teachers, for example, mentioned that the written teacher assessment can be a valuable tool for communication with the parents during parent meetings but also generally throughout the schoolyear.

Effort needed

This category covers benefits and challenges related to the effort needed for successful formative assessment. Two sub-categories emerged from the teachers’ quotes: Time and practice.

Time refers to the preparation time in case of written teacher assessment; lesson time in case of self- and peer-assessment; and so on.

Practice refers to the practice needed by the students so that the formative assessment methods may yield an effect on their learning.

Frequency of the benefits and challenges mentioned

As pointed out earlier, the teachers mentioned the same themes when speaking about benefits and challenges of the different assessment methods. However, some themes were more often referred to when speaking about the benefits, other themes when speaking about the difficulties related to the different methods of assessment. The frequency of these categories mentioned was analysed (independent of the assessment method).

The respective result for the benefits can be found in Figure 29. The results show that all categories apart from 'embedding formal formative assessment in the context of inquiry-based science education', 'use of the feedback', and 'relation between formative and summative assessment' were mentioned from teachers at both school levels. At primary school, the 'content of the feedback' as well as the 'learning effects' were mentioned most frequently. At upper secondary school, the 'learning effects' were mentioned most frequently.

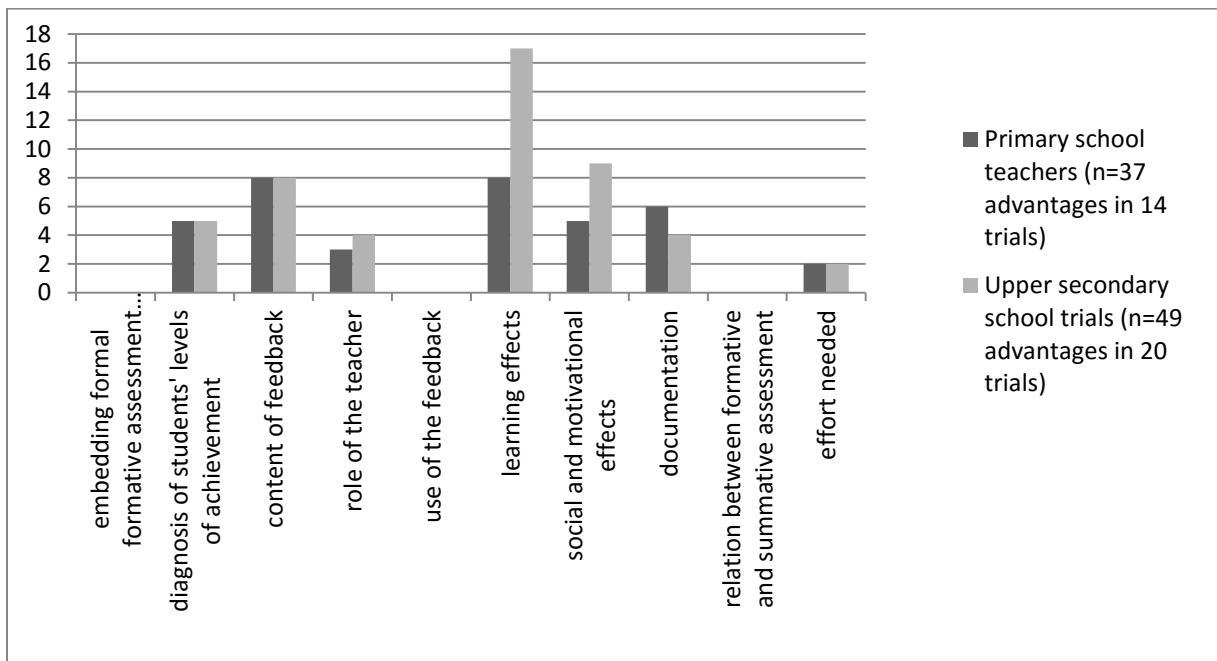


Figure 29: Benefit sub-categories as mentioned by the teachers in the study.

The results for the challenge categories can be found in Figure 30. Apart from 'learning effects', all challenges were mentioned by teachers from at least one school level, with the challenges associated with the 'diagnosis of the student levels', 'the content of the feedback' as well as the 'effort needed' being the most frequently mentioned by both the primary school teachers and the upper secondary school teachers.

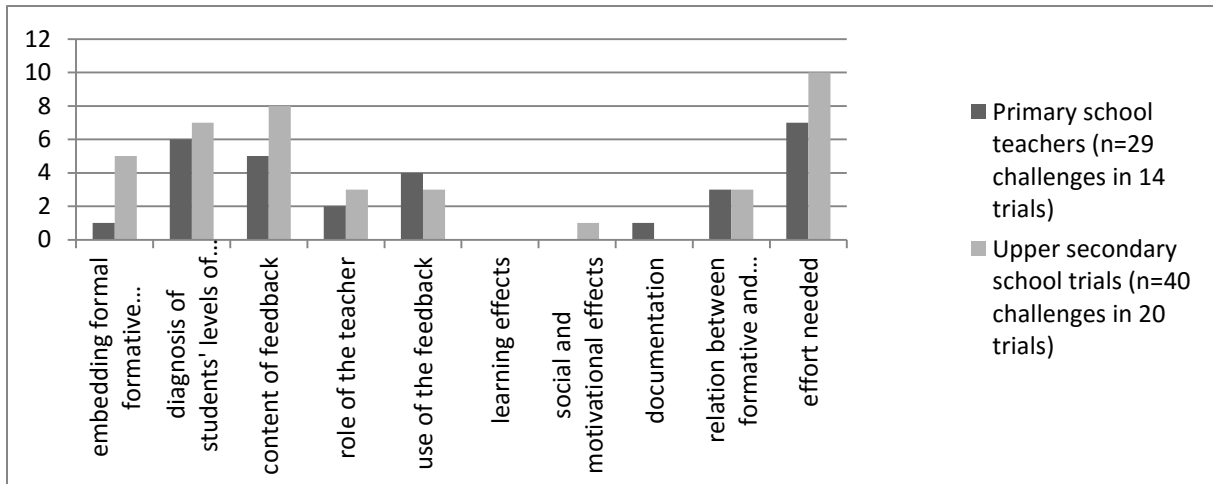


Figure 30: Challenge sub-categories mentioned by the teachers in the study.

Benefits and challenges of written teacher assessment

The categories were generally introduced in the previous sections. Now, the benefits and challenges mentioned by the teachers when speaking about the different methods of formative assessment will be laid out.

The benefits of written teacher assessment as a method of formative assessment (n=17 trials; from which 8 are from primary, 9 are from upper secondary school) covered the following categories: Diagnosis of students' levels of achievement; content of the feedback; learning effects; social and motivational effects; and documentation. The advantages mentioned by the teachers were paraphrased and allocated to the respective sub-categories as introduced in

Table 27.

Table 27: Benefits of written teacher assessment (n=17 trials from 12 teachers): Categories and sub-categories (left) as well as paraphrases of quotes (right).

Categories and sub-categories		Primary school trials (n=8 trials from 6 teachers)	Upper secondary school trials (n=9 trials from 6 teachers)
Diagnosis	Time pressure	- Teacher can take his/her time in deciding about the comments, in setting the priorities	- Teacher can take his/her time in deciding about the comments, in setting the priorities
	Pre-defined criteria	- Diagnosis creates transparency in terms of the expectations for the final grading - Diagnosis can also be provided by a team-teaching partner	- Diagnosis creates transparency in terms of the expectations for the final grading
	Individuality	- Diagnosis has the potential of setting individual standards - Diagnosis is more tightly linked to the people than to their work - Diagnosis allows for more nuanced statement than grade - Diagnosis makes the students' learning process visible	- Diagnosis has the potential of setting individual standards - Diagnosis is more tightly linked to the people than to their work - Diagnosis allows for more nuanced statement than grade
Content of feedback	Potential for enhancement of learning	- Written teacher assessment has the potential of further developing projects and enhancing student-centred activities - Written teacher assessment raises the students' awareness of the learning goals - Written teacher assessment allows for easily drawing conclusions on the further learning	- Written teacher assessment has the potential of further developing projects and enhancing student-centred activities
Learning effects	Scientific concepts	- Written teacher assessment fosters conceptual understanding and content learning	- Written teacher assessment fosters conceptual understanding and content learning
	Transversal competences	- Written teacher assessment fosters personal development	- Written teacher assessment fosters personal development
Social /mot. effects	Relation between teacher and student	- Written teacher assessment improves the student-teacher relation	- Written teacher assessment improves the student-teacher relation
	Motivation	- Written teacher assessment shows the appreciation of the students' work and therefore motivates students	- Written teacher assessment shows the appreciation of the students' work and therefore motivates students
Documentation	Record of feedback for students		- Students cannot simply forget the feedback
	Communication with parents	- Written teacher assessment is documented which makes them a valuable tool for communication with parents	

The challenges of written teacher assessment as a method of formative assessment (n=17 trials; from 8 are from primary, 9 are from upper secondary school) covered the following categories: Embedding of formal formative assessment methods in inquiry-based science education; content of feedback; use of feedback; documentation; relation between formative and summative assessment; and effort needed. The challenges will be introduced with the respective sub-categories in

Table 28.

Table 28: Challenges of written teacher assessment (n=17 trials from 12 teachers): Categories and sub-categories (left) as well as paraphrases of quotes (right).

Categories and sub-categories		Primary school trials (n=8 trials from 6 teachers)	Upper secondary school trials (n=9 trials from 6 teachers)
Embedding	Short-term planning aspects	- Criteria have to be set in the beginning and cannot be changed or adapted during the course of the unit	- Criteria have to be set in the beginning and cannot be changed or adapted during the course of the unit
Content of feedback	Focus	- The quantity of the comments is limited; the teachers have to choose what aspects to concentrate on	- The quantity of the comments is limited; the teachers have to choose what aspects to concentrate on - Support minders the student-centred nature of the learning activities
Use of feedback	Eagerness of recipients	- Some students may not want any feedback	
	Understanding of feedback	- Student understanding of the feedback may differ from teacher understanding	
	Transfer	- Transfer of the feedback to new situations is difficult	- Transfer of the feedback to new situations is difficult
Documentation	Record of feedback for teacher	- Difficult for the teacher to keep the overview over all activities going on	
Relation f. a. to s. a.	Check-like character	- Assessment hinders the joy and the interest in conducting experiments - Criteria for assessment may hinder the openness of inquiries	- Assessment hinders the joy and the interest in conducting experiments - Criteria for assessment may hinder the openness of inquiries
Effort	Time	- Written assessment takes a lot of preparation time for the teacher	- Written assessment takes a lot of preparation time for the teacher

Benefits and challenges of peer-assessment

The benefits of peer-assessment (n=17 trials by 13 teachers; 7 trials from primary school, 10 trials from upper secondary school) covered the following categories: Diagnosis of students' levels of achievement; content of feedback; role of teacher; learning effects; social and motivational effects; and effort needed. The challenges will be introduced with the respective sub-categories in Table 29.

Table 29: Benefits of peer-assessment as mentioned by the teachers (n=17 trials from 13 teachers): Categories and sub-categories (left) as well as paraphrases of quotes (right).

Categories and sub-categories		Primary school teachers (n=7 trials from 5 teachers)	Upper secondary school teachers (n=10 trials from 8 teachers)
Diagnosis	Pre-defined criteria	- Peer-assessment can be provided objectively with criteria	- Peer-assessment can be provided objectively with criteria
Content of feedback	Timing	- Feedback comes immediately	
	Quality in terms of language	- Feedback is easily understandable for students because the language and vocabulary used is familiar	
	Relation between assessor and assessee		- Feedback, particularly criticism, is easier to accept when it comes from peers - Feedback is taken serious since it comes from peers - Inhibition level to ask back in case feedback cannot be understood is low
	Potential for enhancement of learning		- Peer-assessment is for strong and for weak students: weak students receive help, strong students can explain themselves

Table 29 cont.: Benefits of peer-assessment as mentioned by the teachers (n=17 trials from 13 teachers): Categories and sub-categories (left) as well as paraphrases of quotes (right).

Categories and sub-categories		Primary school teachers (n=7 trials from 5 teachers)	Upper secondary school teachers (n=10 trials from 8 teachers)
Role of teacher	Responsibility for student learning		- Students take responsibility for their own learning
	Workload for teacher	- Peer-assessment reduces workload for teacher	- Peer-assessment reduces workload for teacher
	Capacity for individual support	- Peer-assessment provides the teacher with the opportunity to take care of individual difficulties	- Peer-assessment provides the teacher with the opportunity to take care of individual difficulties
Learning effects	Scientific concepts		- Peer-assessment improves engagement with content
	Science-specific competences		- Peer-assessment improves skills / competences assessed
	Transversal competences	- Peer-assessment fosters collaboration in groups; social development - Peer-assessment fosters communication; feedback culture; distinguish between social effects and subject-specific evaluations - Peer-assessment fosters personal development	- Peer-assessment fosters collaboration in groups; social development - Peer-assessment fosters communication; feedback culture; distinguish between social effects and subject-specific evaluations - Peer-assessment fosters personal development - Peer-assessment provides an insight in other students' approaches and solutions which extends personal horizon
	Self-regulated learning	- Peer-assessment fosters self-assessment; reflections	- Peer-assessment fosters self-assessment; reflections
Social and motivational effects	Relation between teacher and student		- Peer-assessment is a way to take students serious and to give value to what they say
	Classroom climate	- Peer-assessment enhances the relation between the students	- Peer-assessment enhances the relation between the students - Peer-assessment is a way for students to show their respect towards other students
	Motivation	- Peer-assessment is liked by the students	- Peer-assessment is liked by the students
Effort	Time	- Peer-assessment needs little time for the teacher to prepare	- Peer-assessment needs little time for the teacher to prepare
	Practice	- Peer-assessment does not need a great lot of introduction	

The challenges of peer-assessment (n=17 trials by 13 teachers; 7 trials from primary school, 10 trials from upper secondary school) covered the following categories: Embedding formal formative assessment methods in inquiry-based science education; diagnosis of students' levels of achievement; content of the feedback; role of the teacher; use of the feedback; social and motivational effects; relation between formative and summative assessment; and effort needed. The challenges will be introduced with the respective sub-categories in Table 30.

Table 30: Challenges related to peer-assessment (n=17 trials from 13 teachers): Categories and sub-categories (left) as well as paraphrases of quotes (right).

Categories and sub-categories		Primary school teachers (n=7 trials from 5 teachers)	Upper secondary school teachers (n=10 trials from 8 teachers)
Embedding	Long-term planning aspects		<ul style="list-style-type: none"> - Peer-assessment activities have to be carefully planned in the course of a semester so that they do not become boring - Peer-assessment is not an assessment method for all students but for social classes mostly
	Short-term planning aspects		<ul style="list-style-type: none"> - Peer-assessment needs students to have their artefacts ready; cannot be postponed to later - Peer-assessment interrupts course of the investigation - Students need different amounts of time to complete this kind of work
Diagnosis	Pre-defined criteria	<ul style="list-style-type: none"> - Students may have difficulties in distinguishing between sympathy and objective criteria - If criteria are not clear, students tend to focus on formal issues rather than on more relevant competences 	<ul style="list-style-type: none"> - Students may have difficulties in distinguishing between sympathy and objective criteria - If criteria are not clear, students tend to focus on formal issues rather than on more relevant competences
	Quality of diagnosis	<ul style="list-style-type: none"> - Students may perceive the peer's level of achievement different than the teacher would 	<ul style="list-style-type: none"> - Students may perceive the peer's level of achievement different than the teacher would - Not all students are equally critical
Content of the feedback	Quality in terms of content	<ul style="list-style-type: none"> - Not all students may take their role serious 	<ul style="list-style-type: none"> - Not all students may take their role serious - Feedback may be unspecific so that it is difficult to derive conclusions from it - The feedback is less reliable in terms of content than the feedback provided by the teacher - Mistakes and misconceptions can be established
	Quality in terms of language and vocabulary	<ul style="list-style-type: none"> - Feedback may be formulated in a way it is hard to draw conclusions from for further learning - Students may not be familiar with feedback rules 	<ul style="list-style-type: none"> - Feedback may be formulated in a way it is hard to draw conclusions from for further learning
Role of teacher	Responsibility for student learning	<ul style="list-style-type: none"> - Teacher cannot check all feedback 	<ul style="list-style-type: none"> - Teacher cannot check all feedback - When to interfere if student coaches oversee mistakes?
Use of feedback	Eagerness of the recipients	<ul style="list-style-type: none"> - Engagement with the feedback depends on how critical and eager the individual students are 	<ul style="list-style-type: none"> - Engagement with the feedback depends on how critical and eager the individual students are - Feedback from peers is often applied without critical reflection on its validity
Social/mot. effects	Motivation		<ul style="list-style-type: none"> - Peer-assessment is rather boring if all students have the same solution
Relation f.a. – s.a.	Relevance of formative assessment	<ul style="list-style-type: none"> - Peer-assessment is not reliable for summative assessment 	<ul style="list-style-type: none"> - Peer-assessment is not reliable for summative assessment
Effort	Time	<ul style="list-style-type: none"> - Peer-assessment takes a lot of time in the lesson 	<ul style="list-style-type: none"> - Peer-assessment takes a lot of time in the lesson
	Practice	<ul style="list-style-type: none"> - Peer-assessment needs training 	<ul style="list-style-type: none"> - Peer-assessment needs training

Benefits and challenges of self-assessment

The benefits of self-assessment (n=6 trials from 6 different teachers; 3 trials from primary school, 3 trials from upper secondary school) covered the following categories: Role of the teacher; learning effects; and social and motivational effects. They will be introduced in detail in Table 31 below.

Table 31: Benefits of self-assessment as perceived by the teachers (n=6 trials from 6 teachers): Categories and sub-categories (left) as well as paraphrases of quotes (right).

Categories and sub-categories		Primary school teachers (n=3 trials from 3 different teachers)	Upper secondary school teachers (n=3 trials from 3 different teachers)
Role of teacher	Capacity for individual support		<ul style="list-style-type: none"> - Self-assessment provides teachers with an insight in students' way of working and thinking - Self-assessment supports teachers in identifying mistakes that occur again and again
Learning effects	Transversal competences	<ul style="list-style-type: none"> - Self-assessment fosters students' abilities to express their opinion and communication skills - Self-assessment fosters students' social abilities 	
	Self-regulated learning	<ul style="list-style-type: none"> - Self-assessment fosters students' autonomy as learners 	<ul style="list-style-type: none"> - Self-assessment fosters students' autonomy as learners
Social/mot. effects	Relation between teacher and student		<ul style="list-style-type: none"> - Self-assessment enhances the relation to students

The challenges of self-assessment (n=6 trials from 6 different teachers; 3 trials from primary school, 3 trials from upper secondary school) covered the following categories: Diagnosis; role of the teacher; and effort needed. They will be introduced in detail in Table 32 below.

Table 32: Challenges of self-assessment as perceived by the teachers (n=6 trials from 6 teachers): Categories and sub-categories (left) as well as paraphrases of quotes (right).

Categories and sub-categories		Primary school teachers (n=3 trials from 3 different teachers)	Upper secondary school teachers (n=3 trials from 3 different teachers)
Diagnosis	Quality of diagnosis	<ul style="list-style-type: none"> - Students are not always honest to themselves - Quality of the self-assessment depends on the quality of the group (who is a member of the group) 	<ul style="list-style-type: none"> - Students are not always honest to themselves
Role of teacher	Responsibility for student learning	<ul style="list-style-type: none"> - Teacher cannot check all reflections - Self-assessment is difficult for the teacher since it is hard to decide when she/he should interfere 	<ul style="list-style-type: none"> - Teacher cannot check all reflections
Effort	Time	<ul style="list-style-type: none"> - Self-assessment consumes a lot of lesson time 	<ul style="list-style-type: none"> - Self-assessment consumes a lot of lesson time
	Practice	<ul style="list-style-type: none"> - Self-assessment needs practice in order to become productive 	<ul style="list-style-type: none"> - Self-assessment needs practice in order to become productive

In the study, three formative assessment methods were trialled. From the teachers' evaluations of the benefits and the challenges of the different assessment methods, ten categories, covering the quotes for both school levels, emerged: Embedding formal formative assessment methods in inquiry-based science educa-

tion; diagnosis of students' levels of achievement; content of feedback; role of the teacher; use of the feedback; learning effects; social and motivational effects; documentation; relation between formative and summative assessment; effort needed. Some of the categories were used to code either benefits or challenges, others were used for both.

The frequency of the different categories mentioned across methods was analysed. Considering the benefits, the content of the feedback as well as the learning effects were mentioned most frequently at primary school. At upper secondary school, the learning effects were mentioned most frequently. Considering the challenges, the diagnosis of the student levels and the content of the feedback as well as the effort needed being the most frequently mentioned by the primary school teachers and the same sub-categories plus the challenges associated with the embedding of formal formative assessment being the most frequently mentioned by the upper secondary school teachers.

Speaking about written teacher assessment, the teachers from both school levels mentioned the advantages related to the quality of the diagnosis and the respective feedback to the students which was expected to lead to student learning in terms of scientific concepts and transversal competences, but also to an effect in the relation between teachers and students and an effect on student motivation. On the other hand, the definition of assessment criteria beforehand, the limited amount and extension of feedback, the doubtful use of the feedback by the students, the check-like character of written teacher assessment, and the big effort in terms of time were mentioned as challenges of written teacher assessment.

On peer-assessment, the teachers in the study mentioned the following advantages: The quality of the feedback in terms of language and its acceptance due to the fact that the assessor is a peer; the responsibility for the learning which lies with the students, resulting in a lower workload for the teacher and a higher capacity for individual support; learning effects in terms of transversal competences (communication etc.) as well as effects on the classroom climate and the students' motivation. Lastly, the low preparation time for the teacher was mentioned. Considering the challenges, the teachers from upper secondary school mentioned difficulties related to the planning of peer-assessment activities. Furthermore, teachers from both school levels expressed their doubts about the quality of the diagnosis and the feedback provided by peers and their uncertainty about their own role. Peer-assessment was also considered rather time-intensive and dependent on a good training of the students.

Speaking about self-assessment, the teachers stressed the advantages related to the role of the teacher who has time for individual support while the students assess themselves. Furthermore, effects in the students' transversal competences and in their self-regulated learning were anticipated. Considering the challenges, the teachers uttered their uncertainty on the quality of the students' reflections and the time such reflections take, similarly to the peer-assessment method.

7.3.3 Means of support for formative assessment as mentioned by the teachers

Apart from reflecting on the benefits and challenges of different formative assessment methods, the teachers in the study were also asked about supportive measures that could facilitate the uptake of such methods in everyday teaching. As in the preceding research sections 7.3.1 and 7.3.2, only the cases which matched the criteria as laid out in sub-chapter 5.4 were included in the analysis. Two data sources were used to explore the question: On the one hand, the teachers were asked about support measures in the evaluation form (n=34 evaluation forms from 21 teachers). On the other hand, the teachers who participated in an individual interview (n=16 interviews with 14 teachers) were also asked the question. The teachers from both school levels suggested almost identical sets of support measures. Therefore, these support measures will not be grouped by school level but introduced overall.

The teachers from primary and from upper secondary school level participating in the study were asked about possible means of support for formative assessment in their classrooms. The teachers' answers were similar across school levels and will therefore be summarized without reference to those. In total, seven

means of supports were mentioned: Provision of examples of good practice; time (e.g. for planning the activities; for providing feedback; etc.); support from team-teaching partner or another person with a teacher-like function; training and coaching to enhance the teachers' assessment literacy; opportunities and prompts to reflect upon assessment practices; platform to exchange experiences and problems with peer teachers; and clarification of the role of formative assessment and its relation to summative assessment at the level of educational policy.

7.3.4 Usability of peer-assessment as mentioned by the students

The successful trial of a formative assessment method in a classroom does not only depend on the teacher but also on the students. It therefore appeared reasonable to also capture the perspective of the students on selected aspects of formative assessment. The data collection was limited with respect to two criteria: Firstly, the data collection on the student perspective was limited to a method where students have a high degree of involvement: Peer-assessment. For the implementation of this method, it appeared particularly useful to explore challenges and suggestions for scaffolding from a student perspective. Secondly, the data collection was limited to a small number of students from only one school level: 5 classes with 103 students in total were asked to fill out a student evaluation form on how they perceived peer-assessment after a respective trial (for details see section 5.3.8). The classes were all from upper secondary school due to the means of data collection (written).

In the above-mentioned evaluation from, the students were asked to what extent they considered peer-assessment useful and why. Table 33 displays the usability of writing feedback to peers as evaluated by the students. The students were generally positive both about assessing: Over 80% of the students in all three settings questioned perceived the role as rather or very valuable.

Table 33: Usability of writing comments in the context of peer-assessment as evaluated by the students. 4=very useful; 1=very useless. AM = arithmetic mean; SD = standard deviation. Classes 1-3 are summarized as one setting because the respective peer-assessment was conducted by the same teacher in the same inquiry-based unit with three different classes.

Usability of writing peer-assessment	4	3	2	1	AM	SD
Usability of writing peer-assessment, classes 1-3 (n=63)	7	46	6	4	2.88	0.69
Usability of writing peer-assessment, class 4 (n=19)	2	14	3		2.95	0.51
Usability of writing peer-assessment, class 5 (n=21)	1	17	3		2.90	0.43

The students were also asked to what extent they considered the feedback they received from peers useful. Table 34 displays the results. The students were generally positive both about receiving feedback: Over 70% of the students questioned perceived the role as rather or very valuable.

Table 34: Usability of comments from peers as evaluated by the students. 4=very useful; 1=very useless. AM = arithmetic mean; SD = standard deviation. Classes 1-3 are summarized as one setting because the respective peer-assessment was conducted by the same teacher in the same inquiry-based unit with three different classes.

Usability of receiving peer-assessment	4	3	2	1	AM	SD
Usability of receiving peer-assessment, classes 1-3 (n=63)	6	42	8	7	2.75	0.78
Usability of receiving peer-assessment, class 4 (n=19)	4	10	3	2	2.84	0.87
Usability of receiving peer-assessment, class 5 (n=21)	7	13	1		3.29	0.55

Selected classes with upper secondary school were asked about the usability of peer-assessment from their perspective. Generally, the students perceived both the role as assessors and assesses positively. Reasons for these evaluations can be found in the subsequent sections.

7.3.5 Benefits and challenges of peer-assessment as mentioned by the students

In this section, the benefits and challenges of peer-assessment as mentioned by the students will be introduced. The themes that emerged from the students' evaluations could be completely covered with the categories derived from the respective teacher answers (see 7.3.2). The sub-categories were slightly adapted. The coding system will be introduced in Table 35.

Since formatively assessing peers and receiving peer-assessment were anticipated to be two different processes from the perspective of students, these potentially two processes were investigated separately in the student evaluation form (see section 5.3.8 and appendix A4 for details), and consequently, the answers will also be displayed separately. This is why benefits and challenges of both providing and receiving peer-assessment will be introduced in this section.

Table 35: Coding system for benefits and challenges of peer-assessment (both assessor and assessee roles) as perceived by the students.

Categories (sub-group of the teacher categories from 7.3.2)	Sub-categories (adapted from the teacher sub-categories in 7.3.2)
diagnosis of students' levels of achievement	quality of diagnosis
content of feedback	focus quality in terms of content quality in terms of language and vocabulary relation between assessor and assessee
learning effects	scientific concepts nature of science science-specific competences transversal competences self-regulated learning other
social and motivational effects	classroom climate motivation
effort needed	time

Benefits of peer-assessment as mentioned by the students

The benefits of assessing peers as mentioned by the students (n=103 students from 5 classes) fell into 2 categories. These will be introduced with the respective sub-categories in Table 36 below. The categories are 'learning effects' and 'social and motivational effects'.

Table 36: Benefits of assessing peers as reported by the students (n=103 students from 5 classes). Categories and sub-categories (left) as well as paraphrases of quotes (right).

Categories and sub-categories		Benefits of assessing peers (n=103 students from 5 classes)
Learning effects	Scientific concepts	- Peer-assessment fosters content knowledge; the revision and engagement with the content
	Nature of science	- Peer-assessment fosters students' understanding of why it is important to describe and explain exactly; to write precise protocols; to structure protocols properly; to label sketches - Peer-assessment fosters students' understanding of why more than one problem solving strategy and more than one solution may be suitable for one task
	Science-specific competences	- Peer-assessment provides opportunity to get insight to other problem solving strategies and new ideas
	Transversal competences	- Peer-assessment fosters students' communication abilities: to write specific feedback; to praise good aspects of the work; to bring up critical points; to assess along criteria
	Self-regulated learning	- Peer-assessment provides opportunity to compare own artefact to the artefact of others - Peer-assessment provides opportunity to see peers' mistakes and be sensitized to avoid them oneself - Peer-assessment fosters students' self-reflection abilities and a critical view of their own work - Peer-assessment fosters students' abilities to carry the responsibilities of their own and their peers' learning
Social and motivational effects	Classroom climate	- Peer-assessment provides the opportunity to help others
	Motivation	- Peer-assessment is a nice variation from ordinary lessons

The benefits of receiving feedback from peers as mentioned by the students (n=103 students from 5 classes) fell into 3 categories. These will be introduced with the respective sub-categories in Table 37 below. The categories are 'content of the feedback', 'learning effects' and 'social and motivational effects'.

Table 37: Benefits of receiving feedback from peers as reported by the students (n=103 students from 5 classes). Categories and sub-categories (left) as well as paraphrases of quotes (right).

Categories and sub-categories		Benefits of receiving feedback from peers (n=103 students from 5 classes)
Content of feedback	Focus	- Peer-assessment focusses more on details whereas teacher assessment focusses more on the overall picture - Feedback from peers is easier to understand than teacher assessment (precisely linked to a detail and not general) - Feedback from peers points to different aspects compared to feedback from the teacher
	Relation between assessor and assessee	- Feedback from peers is considered more than feedback from the teacher - Feedback from peers is more personal than feedback from the teacher

Table 37 cont.: Benefits of receiving feedback from peers as reported by the students (n=103 students from 5 classes). Categories and sub-categories (left) as well as paraphrases of quotes (right).

Categories and sub-categories		Benefits of receiving feedback from peers (n=103 students from 5 classes)
Learning effects	Scientific concepts	- Peer-assessment fosters content knowledge; the revision and engagement with the content
	Science-specific competences	- Peer-assessment fosters the understandability of formulation and sketches: because the teacher usually understands what students tried to express anyways but the peers do not necessarily
	Transversal competences	- Peer-assessment fosters abilities on formal and layout-related abilities
	Self-regulated learning	- Peer-assessment fosters students' self-reflection abilities and a critical view of their own work
Social and motivational aspects	motivation	- Peer-assessment is motivating due to the motivating comments - Peer-assessment is less critical than teacher assessment and therefore enhances self-confidence

Challenges of peer-assessment as mentioned by the students

The challenges of assessing peers as mentioned by the students (n=103 students from 5 classes) fell into four categories. These will be introduced with the respective sub-categories in Table 38 below. The categories are 'content of feedback', 'learning effects', 'social and motivational effects', and 'effort needed'.

Table 38: Challenges of assessing peers as reported by the students (n=103 students from 5 classes). Categories and sub-categories (left) as well as paraphrases of quotes (right).

Categories and sub-categories		Challenges of assessing peers (n=103 students from 5 classes)
Content of feedback	Quality in terms of content	- Peer-assessment can be subjectively influenced (e.g. too kind because the feedback is directed to a peer) - Assessors are uncertain about what feedback to provide
Learning effects	Other	- Assessors can only correct mistakes they already are aware of; they are therefore unable to improve themselves - Assessors are influenced by the artefacts they have to assess, they may copy mistakes
Social and motivational effects	Motivation	- Peer-assessing (criticising and correcting) is an unsympathetic role - Peer-assessing is boring
Effort needed	time	- time

The challenges of receiving feedback from peers as mentioned by the students (n=103 students from 5 classes) fell into three categories. These will be introduced with the respective sub-categories in Table 39 below. The categories are 'diagnosis of students' levels of achievement', 'content of feedback', and 'social and motivational effects'.

Table 39: Challenges of receiving feedback from peers as reported by the students (n=103 students from 5 classes). Categories and sub-categories (left) as well as paraphrases of quotes (right).

Categories and sub-categories		Challenges of receiving feedback from peers (n=103 students from 5 classes)
Diagnosis of students' levels of achievement	Quality of diagnosis	<ul style="list-style-type: none"> - Peers have different levels of achievement and therefore provide feedback of variable quality - Peers have variable "levels of tolerance" before they comment on a mistake or on an imprecise formulation
Content of feedback	Focus	<ul style="list-style-type: none"> - Peer-assessment often focusses on details which are not so important (e.g. formal issues)
	Quality in terms of content	<ul style="list-style-type: none"> - Peers do not have the necessary background in terms of content knowledge; they may not recognize mistakes or even pass on their misconceptions - As an assessee, one cannot be sure about the quality of the feedback received
	Quality in terms of language and vocabulary	<ul style="list-style-type: none"> - Peer-assessment is sometimes formulated in a way it is difficult to derive next steps of learning from it: little details, low level of concreteness, rather general ("good"), superficial - Peer-assessment is often not structured well or language-wise hard to understand
Social and motivational effects	Classroom climate	<ul style="list-style-type: none"> - Peer-assessment can provide compliments that are not justified - Peer-assessment can be hurtful

In the study, students from five upper secondary school classes were asked about the benefits and challenges of peer-assessment they perceived. The themes that emerged from the students' evaluations could be completely covered with a sub-set of the categories derived from the respective teacher answers. Speaking about the benefits, the students mentioned 'content of feedback', 'learning effects', and 'social and motivational effects'. Speaking about the challenges, the students mentioned 'diagnosis of students' levels of achievement', 'content of feedback', 'social and motivational effects', and 'effort needed'.

7.3.6 Means of support for peer-assessment as suggested by the students

This section will introduce means of support that could scaffold peer-assessment as suggested by students. The question was included in the student evaluation form (see section 5.3.8) and was answered by n=103 students from 5 classes. The answers will be summarized below.

Some students said that no support was needed; that they could peer-assess *"just like that"*. Other students addressed the difficulty of providing feedback. In order to support this communication process, they suggested examples of good practice on what good peer-assessment should look like. One of the students put it like that: *"I didn't know what comments and suggestions would be constructive in the beginning. So a good example would help."* Similarly, some students asked for structuring questions or criteria that provide guidance in what should be focussed on. Furthermore, it was suggested that the feedback should be provided anonymously. Other students addressed their uncertainty due to the lack of background knowledge. They suggested that access to detailed content knowledge would help. Adding to this issue, one of the students wrote: *"It is important that the assessor has successfully completed the task he/she is assessing."* A similar idea was to provide access to a correct solution or an example of a very good solution. Some students suggested that a platform with all solutions that everyone could check would help. Furthermore, the students asked for a possibility to exchange with peers or even with the teacher.

In the study, students from five upper secondary school classes were asked about possible means of support for formative peer-assessment in their classrooms. The students' answers fell into six categories: (1) no support needed; (2) support in formulating feedback; (3) structuring questions or criteria to focus on; (4)

anonymity; (5) access to content knowledge or to the correct solution; and (6) exchange with peers or with the teacher.

7.4 Results on research question 4: Changes in teachers' understandings and implementations throughout the collaboration in the study

In the fourth part, research question 4: *Changes in the teachers' understandings and implementations of formative assessment throughout the collaboration in the study* will be investigated. The aim of this sub-chapter is to search for possible changes in the teachers' understanding of formative assessment (section 7.4.1); or in their self-efficacy (section 7.4.2); changes in the teachers' formative assessment practices (section 7.4.3); changes in their perceptions of importance, benefits and challenges related to formative assessment (section 7.4.4); and the support mechanisms from the collaboration in the study as mentioned by the teachers (section 7.4.5). The last section (7.4.6) summarizes the teachers' implementation behaviours as represented in the variability of their trials. The data is the same which had already been coded for research questions 1, 2, and 3 but is this time analysed dependent on the round of implementation. Additionally, the teacher profile questionnaire (see section 5.3.1) will be analysed using non-parametric tests.

With the small sample sizes, the results on research question 4 are clearly tenuous. Due to the little literature available on changes in teachers' formative assessment practices and beliefs throughout the collaboration in a project where the teachers develop their own assessment, it nevertheless appeared legitimate to conduct the respective analyses. The results will be interpreted with caution. Part of this cautious interpretation is that the data will, in some sections, not be analysed separately for the two school levels as for the other research questions. Instead, the teachers from the two school levels will be considered as one group.

7.4.1 Changes in the teachers' descriptions of what formative assessment is throughout the study

In order to get some idea about how the teachers' understandings of formative assessment changed throughout the course of the study, all definitions of teachers whose data were available for at least two measurement points were analysed over time. The results can be found in Table 40. The detailed description of the codes can be found in sub-chapter 7.1. As in the analysis for research question 1, it was distinguished between elements ascribed to formative assessment in the literature (a-e) and other elements (g-j). Elements f) 'criterion-based' and k) 'examples of assessment methods' were excluded: In the case of 'criterion-based', this characteristic cannot be ascribed to formative assessment only but to assessment in general. In the case of the latter, being able to list the examples of assessment methods from the study was not considered a clear sign of a conception.

The results show that in the first semester of collaboration, there were four teachers from both school levels who mentioned only elements which are also ascribed to formative assessment in the literature. Another four teachers, also coming from both school levels, mentioned elements that are not ascribed to formative assessment in the literature exclusively. These elements represented two basic ideas: Firstly that formative assessment is grading of the learning process and secondly unclear references or explications on inquiry-based education. Finally, four teachers mentioned both elements which are ascribed to formative assessment in the literature but also other elements.

In the second semester of collaboration, nine teachers from both school levels mentioned only elements that are also ascribed to formative assessment in the literature in their definitions. One teacher from upper secondary school mentioned exclusively elements that are not ascribed to formative assessment in the literature. Instead, this teacher made unclear and inquiry-related references. Four teachers from both school levels mixed elements that are ascribed to formative assessment in the literature and elements that are not. The later included three ideas: That formative assessment focusses on a specific set of competences such as social behaviour or organisational skill; that formative assessment is individual-referenced; and unclear and inquiry-related references.

Table 40: Elements in the teachers' written definitions in the three rounds of implementation. Since the teachers' were asked about their understanding of the term 'formative assessment' anonymously, their codes are different from the codes introduced in sub-chapter 5.2.

		1 st semester of collaboration					2 nd semester of collaboration					3 rd semester of collaboration				
		Elements from literature	Other elements				Elements from literature	Other elements				Elements from literature	Other elements			
		a)-e)	g)	h)	i)	j)	a)-e)	g)	h)	i)	j)	a)-e)	g)	h)	i)	j)
Primary school teachers	AW28				x	x	x		x			x				x
	MH32	x					x					x				
	MS35					x	x									---
	VO40	---					x					x				x
	VZ31	x				x	x					x				x
	ZF56	x				x	x					x				
Upper secondary school teachers	AF33	x				x	x	x			x					x
	AM50	x					x				x					
	DB33	x					x				x					
	FR50					x	x	---				x	x			x
	LS27						x									x
	JS47	---					x					x				
	RG53	---					x	x			x	x				x
	TB21	x					x					x				
	UP42	x					x					x				

In the third semester of collaboration, seven teachers from both school levels mentioned only elements that are also ascribed to formative assessment in the literature in their definitions. One teacher (the same as in round two) mentioned exclusively elements which are not ascribed to formative assessment in the literature. This teacher made unclear and inquiry-related references as in the second round. Six teachers from both school levels mentioned both elements that are ascribed to formative assessment in the literature and elements that are not. These other elements included, on the one hand, the understanding that formative assessment focusses on a specific set of competences. On the other hand, the other elements included unclear and inquiry-related references.

Overall, the number of definitions which are completely consistent with the literature on formative assessment increased from the first semester (four cases) to the second semester of collaboration (nine cases). The number decreased again in the third semester (seven cases). The number of definitions that do not contain any elements that were ascribed to formative assessment in the literature decrease from the first semester (four cases) to the later semesters of collaboration (the same one case in ssemesters two and three).

Considering the changes between the rounds of implementation (first, second, third semester) at the level of individual teachers, four groups can be made: In the first group are the teachers whose understanding of

formative assessment neither improved nor worsened; e.g. the teachers who explained what formative assessment is with elements that are also present in the literature (seven teachers). In the second group are the teachers whose understanding of formative assessment improved; e.g. from explanations that included both elements that are consistent to the literature and others towards only elements that are also mentioned in the literature (five teachers). In the third group are teachers whose understanding of what formative assessment worsened throughout the collaboration in the study; e.g. from elements that are also mentioned in the literature to both elements from the literature and others (one teacher). In the fourth group are the two teachers whose development does not go into a clear direction (MH32, VZ31).

Considering the elements that were mentioned in the teachers' definitions but are not ascribed to formative assessment in the literature, there are considerable changes between the rounds: The idea that formative assessment was grading of the learning process was only expressed in the first round (four teachers from both school levels). In the second and in the third round, the idea that formative assessment must focus on a specific set of competences such as social skills appeared (three teachers at upper secondary school level). Only in the second round, the idea that formative assessment is individual-references was expressed (one teacher at primary school level). Unclear and inquiry-related references were made in all rounds.

The length of the explanations did not vary a lot: on average, the definitions contained 50 words in the first round, 47 words in the second round and 55 words in the third round of implementation.

The teachers' descriptions of what formative assessment and their changes throughout the collaboration (beginning of collaboration; middle phase; end of collaboration) in the study were analysed. The 11 categories developed to code the teachers' answers for research question 1 were used again: Five categories describe elements that are also present in the literature on formative assessment; one element is generally ascribed to assessment in the literature; four categories describe other elements and one category contains examples of assessment methods. The general analysis shows that the teachers' descriptions of formative assessment generally converged towards what can also be found in the literature (supportive nature; guidance for next steps in learning or teaching; individual; prospective rather than retrospective). The analysis throughout the time of collaboration shows that four groups of teachers can be made: The first group in which the teachers' understanding of formative assessment did not change throughout the collaboration in the study (seven teachers); the second group in which the teachers improved towards descriptions which can also be found in the literature (five teachers); the third group in which the teachers' descriptions worsened (the opposite direction than the second group; one teacher); and unclear directions (two teachers).

7.4.2 Changes in the teachers' self-efficacy throughout the collaboration in the study

In the teacher profile questionnaire, the teachers who collaborated in the study and a control group were asked about their personal formative assessment self-efficacy belief (see section 5.3.1 for the description of the teacher profile questionnaire; see section 5.5.4 for the introduction of the personal formative assessment efficacy belief scale). The aim was to explore changes between the teachers' answers from Sept 2014 (first measurement, beginning of the project for the teachers) and January 2016 (second measurement, end of the project). The descriptive statistics of the respective scale can be found in Table 41.

Table 41: Descriptive statistics of the personal formative assessment efficacy belief scale and its items. Mdn=median; AM= arithmetic mean; SD= standard deviation.

	Study teachers (n ₁ =16)						Control group (n ₂ =13)					
	Sept 2014			Jan 2016			Sept 2014			Jan 2016		
	Mdn	AM	SD	Mdn	AM	SD	Mdn	AM	SD	Mdn	AM	SD
Scale: Personal formative assessment efficacy belief (see 5.5.4)	2.69	2.64	0.95	2.00	2.08	0.64	2.75	2.98	1.14	2.67	2.91	1.01

From the differences in the medians of the two study teacher measurements, it can be hypothesized that significant changes in the study teachers' personal formative assessment efficacy belief occurred between

the first and the second measurement. Changes were tested for both the teachers involved in the study and also a control group consisting of science teachers from the same schools and the same school levels. The results of the non-parametric Wilcoxon test and the effect size are displayed in Table 42.

Table 42: Wilcoxon tests and effect sizes for the personal formative assessment efficacy belief scale and its' items. Significance: * $p < 0.05$; ** $p < 0.01$; Cohen's d : $|0.2| < d \leq |0.5|$ represents small effect size; $|0.5| < d \leq |0.8|$ represents medium effect size; $|0.8| < d$ represents large effect size (Cohen, 1988).

	Study teachers ($n_1=16$)			Control group ($n_2=13$)		
	Asymptotic Wilcoxon test		Effect size	Asymptotic Wilcoxon test		Effect size
	z-value	p (2tailed)	Cohen's d	z-value	p (2tailed)	Cohen's d
Scale: Personal formative assessment efficacy belief (see 5.5.4)	-2.45	$p < 0.05$	-0.69	-1.10	n.s.	-0.06

The findings in the scale on personal formative assessment efficacy belief show that the teachers who participated in the study improved significantly from the beginning of the collaboration (median 2.69, low values stand for high personal formative assessment efficacy belief) to the end (median 2.00; asymptotic Wilcoxon-test: $z = -2.45$, $p < 0.05$, $n = 16$) whereas the teachers from the control group did not change significantly. The Cohen effect size of the change of the teachers collaborating in the study is $d = |0.69|$ which represents a medium size effect.

The change in the teachers' personal formative assessment efficacy belief was measured by the respective items in the teacher profile questionnaire at the beginning and at the end of collaboration in the study. The results show that the collaborating teachers' personal formative assessment efficacy belief increased significantly whereas the control groups' personal formative assessment efficacy belief did not change. The control group consisted of peer teachers who did not collaborate in the study.

7.4.3 Changes in the trials throughout the collaboration in the study

Changes in the quality of the formative assessment activities

In sub-chapter 5.4, the selection of cases for analysis has been introduced. It was conducted based on four groups of criteria:

- C1: The conduction of trials
- C2: The documentation of trials
- C3: Criteria related to inquiry units
- C4: Criteria related to planned-for-interaction formative assessment.

In order to get some insight into the changes in the quality of the teachers' implementation, the coded trials of all teachers are displayed over time in Table 43. The results show that in the first semester of collaboration (round), eight teachers from primary school teachers and five teachers from upper secondary school trialled a formative assessment method in their inquiry teaching. One trial from primary school did not match all the criteria as defined in chapter 5.4: The beginning of a formative assessment cycle was present but the students had no opportunity to use the feedback they received. At upper secondary school, two teachers did not implement anything and one teacher did not provide the students with the opportunity to use the feedback they received.

In the second semester of collaboration (round), four teachers from primary school and seven teachers from upper secondary school conducted a trial that matched the criteria as defined in chapter 5.4. Three teachers from primary school enacted a trial but with no formative assessment included: In one of the trials, the criteria for formative assessment were not communicated in advance (or this could not be seen from the documentation). In one case, there was no formative assessment at all. In the third case, the assessment method was not trialled in the context of inquiry-based education. At upper secondary school, two teachers

had trials that did not match all criteria as defined in chapter 5.4. One trial did not take place in the context of inquiry whereas the other did not involve any formative assessment at all.

Table 43: The teachers' implementations displayed over time. Criteria 1) – 4) spelled out above.

Y	Trial matching criteria 1-4
N	Trial not in the context of inquiry or not involving formative assessment (violating criteria C3 or C4) Details on criterion C4 related to planned-for interaction formative assessment: a) Assessment criteria not communicated in advance b) No use of feedback received c) No formative assessment at all
N	No trial or trial not sufficiently documented (violating criteria C1 or C2)
--	Trials that emerged from collaboration with Olia Tsvitanidou and teachers who left the study before its end

		1 st round of implementation				2 nd round of implementation				3 rd round of implementation			
		C1	C2	C3	C4	C1	C2	C3	C4	C1	C2	C3	C4
Primary school teachers	P1	Y	Y	Y	Y	Y	Y	Y	N ^{a)}	--			
	P2	Y	Y	Y	Y	Y	Y	Y	Y	N			
	P3	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
	P4	Y	Y	Y	N ^{b)}	N					N		
	P5	Y	Y	Y	Y	Y	Y	Y	Y	--			
	P6	Y	Y	Y	Y	Y	Y	Y	N ^{c)}		N		
	P7	Y	Y	Y	Y	N				N			
	P8	Y	Y	Y	Y	Y	Y	Y	Y		N		
	P9	Y	Y	Y	Y	Y	Y	N	Y	Y	Y	Y	Y
Upper secondary school teachers	S1	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
	S2	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
	S3	N				Y	Y	Y	Y	Y	Y	Y	Y
	S4	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
	S5	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
	S6	N				Y	Y	Y	Y	N			
	S7	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	N ^{b)}
	S8	--				Y	Y	N	Y	Y	Y	Y	Y
	S9	--				--				Y	Y	Y	N ^{b)}
	S10	--				--				Y	Y	Y	Y
	S11	Y	Y	Y	N ^{b)}	Y	Y	Y	N ^{c)}	Y	Y	Y	Y

In the third round, two teachers from primary school and eight teachers from upper secondary school successfully trialed a formative assessment method in the context of inquiry teaching. At primary school, five teachers either did not trial anything or did not document their trials well enough that it could be analysed. At upper secondary school, two teachers had trials with imperfect formative assessment: In both cases, the students received feedback but did not have the opportunity to use this feedback. One teacher at upper secondary school did not conduct a trial.

Overall, the number of trials that did not take place in the context of inquiry or that did not involve formative assessment is biggest in the second round. The number of cases with no or insufficient documentation seems to increase in the primary school cases (from zero cases in the first round to two cases in the second

and then to five cases in the third round), whereas this same number seems to vary on a small level at upper secondary school (two cases in the first round, no cases in the second round, one case in the third round).

Changes in the quantity of the formative assessment activities

In the teacher profile questionnaire, the teachers who collaborated in the study and a control group were asked about the use of formative and summative assessment in their teaching practice (see section 5.3.1 for the description of the teacher profile questionnaire; see section 5.5.4 for the introduction of the scales). The aim was to explore changes between the teachers' answers from Sept 2014 (first measurement, beginning of the project for the teachers) and January 2016 (second measurement, end of the project). The descriptive statistics of the two respective scales can be found in Table 44. Arithmetic means and standard deviations were not calculated because the scales were not interval-scaled.

Table 44: Descriptive statistics of the frequency scales. Mdn=median.

	Study teachers (n ₁ =16)		Control group (n ₂ =13)	
	Sept 2014	Jan 2016	Sept 2014	Jan 2016
	Mdn	Mdn	Mdn	Mdn
Formative assessment – frequency (see 5.5.4)	3.56	3.89	3.89	4.00
Summative assessment – frequency (see 5.5.4)	4.00	3.33	3.42	4.00

Changes were tested for both the teachers involved in the study and also a control group consisting of science teachers from the same schools and the same school levels. The results of the non-parametric Wilcoxon tests are displayed in Table 45. Effect sizes were not calculated because the scales were not interval-scaled.

Table 45: Wilcoxon tests and effect sizes for the frequency scales. Significance: *p<0.05; **p<0.01.

	Study teachers (n ₁ =16)		Control group (n ₂ =13)	
	Asymptotic Wilcoxon test		Asymptotic Wilcoxon test	
	z-value	p (2tailed)	z-value	p (2tailed)
Formative assessment – frequency scale (see 5.5.4)	-1.11	n.s.	-0.61	n.s.
Summative assessment – frequency scale (see 5.5.4)	-1.19	n.s.	-0.11	n.s.

The results show that neither the teachers directly involved in the study nor their peers changed significantly in their estimation of how often they formatively or summatively assess.

The quality and the quantity of the formative assessment activities trialled throughout the study (in three rounds) were analysed. The resulting picture is not very clear at the level of the individual teacher. However, what can be said is that in the first round, almost all trials conducted followed the criteria as defined in sub-chapter 5.4 and were successful in that sense. In the second round, a considerable number (five trials out of 16) did not fulfil all criteria defined in 5.4. In the third round, most of the primary school teachers either left the study or did not conduct any trials.

In order to analyse the quantity of formative assessment activities, the respective items from the teacher profile questionnaire were analysed. No significant differences between the beginning and the end of the study were measured.

7.4.4 Changes in the importance, benefits and challenges of the formative assessment methods perceived by the teachers throughout the collaboration in the study

Changes in the importance of formative assessment as perceived by the teachers

In the teacher profile questionnaire, the teachers who collaborated in the study and a control group were asked about their perception of the importance of formative and summative assessment (see section 5.3.1 for the description of the teacher profile questionnaire; see section 5.5.4 for the introduction of the scales). The aim was to explore changes between the teachers' answers from Sept 2014 (first measurement, beginning of the project for the teachers) and January 2016 (second measurement, end of the project). The descriptive statistics of the three respective scales can be found in Table 46. Since Cronbach's α was low, no values are displayed for the summative assessment – importance scale.

Table 46: Descriptive statistics of the importance scales. Mdn=median; AM= arithmetic mean; SD= standard deviation.

	Study teachers (n ₁ =16)						Control group (n ₂ =13)					
	Sept 2014			Jan 2016			Sept 2014			Jan 2016		
	Mdn	AM	SD	Mdn	AM	SD	Mdn	AM	SD	Mdn	AM	SD
Formative assessment – importance scale (see 5.5.4)	2.17	2.14	0.57	2.11	2.19	0.53	2.88	2.68	0.48	2.78	2.64	0.61
Summative assessment – importance scale (see 5.5.4)	low Cronbach's α						low Cronbach's α					

Changes were tested for both the teachers involved in the study and also a control group consisting of science teachers from the same schools and the same school levels. The results of the non-parametric Wilcoxon tests are displayed in Table 47.

Table 47: Wilcoxon tests and effect sizes for the importance scales. Significance: *p<0.05; **p<0.01; Cohen's d: |0.2|<d≤|0.5| represents small effect size; |0.5|<d≤|0.8| represents medium effect size; |0.8|<d represents large effect size (Cohen, 1988).

	Study teachers (n ₁ =16)			Control group (n ₂ =13)		
	Asymptotic Wilcoxon test		Effect size	Asymptotic Wilcoxon test		Effect size
	z-value	p (2tailed)	Cohen's d	z-value	p (2tailed)	Cohen's d
Formative assessment – importance scale (see 5.5.4)	-0.41	n.s.	0.09	-0.11	n.s.	-0.07
Summative assessment – importance scale (see 5.5.4)	low Cronbach's α			low Cronbach's α		

In their evaluation of the importance of formative assessment, the results show that neither the teachers directly involved in the study nor their peers from the control group changed significantly. No effect was found with Cohen's d either. The respective scale for inquiry-based education and for summative assessment could not be analysed because the Cronbach's α was low.

Changes in the benefits and challenges of formative assessment as perceived by the teachers

In order to explore possible changes in the in the teachers' perceptions of benefits and challenges of the assessment methods throughout the collaboration in the study, the benefits and challenges as introduced in 7.3 were plotted dependent on the round of implementation, but independent of the assessment method. The later was possible because the categories used to code the benefits and challenges were the same for each assessment method (compare to sub-chapter 7.3).

The results can be found in Table 48. No tendencies or patterns concerning changes throughout the three semesters of implementation could be found.

Table 48: Benefits and challenges perceived by the teachers in the study throughout the three rounds of implementation.

	round 1		round 2		round 3		
	Primary school teachers	Upper sec. teachers	Primary school teachers	Upper sec. teachers	Primary school teachers	Upper sec. teachers	
benefits	diagnosis of students' levels of achievement	2	1	2	3	1	1
	content of feedback	5	1	2	5	1	2
	role of the teacher	2	2	0	1	1	1
	learning effects	5	4	1	5	2	8
	social and motivational effects	3	2	1	3	1	4
	documentation	4	1	2	3	0	0
	effort needed	1	0	0	1	1	1
challenges	embedding formal formative assessment methods in inquiry-based science education	1	2	0	2	0	1
	diagnosis of students' levels of achievement	2	3	3	2	1	2
	content of feedback	2	3	2	2	1	3
	role of the teacher	1	2	0	0	0	1
	use of the feedback	3	0	1	2	0	1
	social and motivational effects	0	0	0	1	0	0
	documentation	1	0	0	0	0	0
	relation between formative and summative assessment	1	1	0	1	2	1
effort needed	4	4	1	2	2	4	

Changes in the teachers' perception of importance of formative assessment and challenges and benefits of the formative assessment methods throughout the study were explored. The results show that the perceived importance of formative assessment did not change. No pattern or tendency could be found in the advantages and challenges mentioned either.

7.4.5 Support mechanisms in the collaboration in the study

In the individual interviews (see section 5.3.9), the teachers from both school levels were asked about the aspect(s) of the collaboration in the study that were most useful to them in terms of their formative assessment practices.

The teachers came up with six basic support mechanisms in the study: Firstly, the background on the theory of formative assessment as provided in the manual, on the dropbox (as texts) but also presented in some of the meetings were mentioned. One of the teachers (P9) said *"like this, I learn about the background of these methods, I can also look up details again later, [...]"*. Another teacher (P4) specifically referred to a graphical representation of the formative assessment cycle in the manual and said *"there, that visualisation is nice, and I can keep track of the sequence I am at <when preparing my units> and what do I want to do now, and what I should concentrate more on. [...]"*.

Secondly, the teachers mentioned the provision of concrete methods for formative assessment along with examples as useful. These methods and also the examples were provided in a written form in the manual but also presented and discussed orally in some of the meetings. One of the teachers (P2) said *"this was a really good refresher <from my pre-service training> and I always thought, ok, that is this method. This would*

also be an option for my teaching". Another teacher (S10) said "I was really unaware of the variety of different methods for formative assessment before it was introduced to me in the project." A third teacher (S11) said "these examples inspire me, and even though I do not like everything in them, I can still do something similar that goes into the same direction. Many of the examples were for primary and lower secondary level but they can be adapted."

A third aspect that was brought up in the individual interviews was the opportunity to try out different methods. A teacher (S2) put it like this: "the chance to try out the methods helps me most. I am just like that. Of course that needs some effort; for preparation and also for the trial itself. I am already thinking about what I could do in the next semester, what would fit the context of a certain topic and class."

The fourth mechanism that supported the formative assessment practices was the reflexions on the own assessment practices that was triggered in various occasions. For one teacher (S10), it was the study as a whole: "the project itself was a big help, firstly to become aware of what formative assessment is, and secondly to become aware of my own assessment practices which actually always included some formative assessment even though I did not know the term [...]". For another teacher (S4), the triggers of reflection were the evaluation forms and the individual interview: "it is a verbalisation of what I did; I have to explain and to evaluate [...]". For a third teacher (S7), the reflections were prompted by the inputs provided during the meetings with all teachers.

Fifthly, the exchange with the other teachers was mentioned as a main support mechanism during the project meetings (two or three meetings per semester). One of the teachers (P9) said that "I get new ideas; sometimes other people have trialled things I could never have thought of myself". Another teacher (S1) said: "what is of course useful are the discussions which we have in our meetings. When we are in the small groups with the other physics teachers and start talking, that is really useful. To see, how do the others do in the lab lessons, how do they coach their students, how do they put an assessment method into practice, what unit would be suitable for this." For this exchange of ideas, the dropbox was explicitly mentioned (S4) as being helpful: "I saw that there are examples on the dropbox [...] and to exchange details and lessons learned, this is beneficial. I will certainly take my materials to the next meeting to tell the others about what I did <in my trial>". Apart from being a source of inspiration, these possibilities for exchange were also a possibility to gain confidence. A teacher (P6) put it like this: "<the exchange with the other teachers> reassured me, you know, even this very experienced colleague said that documentation of formative assessment was a challenge for him. So I thought, ok, I am not the only one who has trouble with this."

The final aspect that was mentioned was the broadening of the horizon through the contact with teachers from a different school level (in this case, with the primary school teachers). One of the teachers (S7) said "[...] just to get to know teachers from other school levels, because the upper secondary school level is not isolated." A second teacher (S10) said "there are also impulses from the project which I cannot use directly in my teaching, but which broaden my horizon. Like, this could also be done, in a different subject. Or that could be done, at primary school level."

The teachers collaborating in the study were asked what aspect(s) of the study they considered helpful: The theoretical background information on formative assessment; the provision of concrete ideas and examples to draw from; the opportunity and the prompts to try out different methods; the opportunity and prompts to reflect upon the formative assessment practices; the exchange with peer teachers; and the broadening of the horizon through the contact with teachers from a different school level.

7.4.6 Variability of implementations within teachers

The extent to which the individual teachers changed their trials from one round to the next round varied. Primary school teacher P1 (see appendix A7), for example, trialled the peer-assessment method in the first semester of collaboration. Her students explored buoyancy by hypothesis generation and testing on what

objects sink and what objects float. The students observed their peers and assessed their hypothesis generation, investigation, analysis and interpretation as well as communication in a rubric with smileys and written advice. This assessment was used to investigate the floating abilities of another, subsequent object. In her second trial, the same teacher trialled a combination of written teacher assessment and peer-assessment. The students worked in groups and observed different animals. The main task was to present the findings on these animals. The teacher herself assessed the students' investigative competence as well as the use of interactive tools, the students assessed their peers' communication. So this teacher made substantial changes to both the inquiry context and the formative assessment from her first to her second trial.

As a contrasting example, upper secondary school teacher S4 (see appendix A7) trialled written teacher assessment in the first semester of collaboration. His students investigated pressure. The main task was to document respective experiments and the teacher assessed the communication of the results based on the students' lab journal entries. The respective feedback was transferred to a subsequent lab unit. In his second trial, the teacher only changed the topic of the inquiry (now electric circuits) and the assessment method (now peer-assessment). The degrees of openness of the inquiry and the activities enacted, the data for diagnosis and the use of feedback remained the same. The changes from the second to the third trial were small again.

Apparently, teachers with different implementation habits collaborated in the study. In the attempt to learn about the implementation histories of the individual teachers, the variability of their implementation was analysed. In more detail, the overlap in the different variables (such as dimensions of openness, inquiry activities etc.) between two subsequent trials of a teacher was calculated from the coding for research question 2. The results can be found in Table 49.

'Overlap' in the context of the different variables (such as dimensions of openness, inquiry activities etc.) describes the size of the intersecting set of options in relation to the total size of options of the same variable. If, for example, a teacher's first trial was coded as open in dimensions of openness A, B, and C, and the same teacher's second trial was coded open in dimensions of openness C and E, the overlap is 25% and will be displayed as 0.25 in Table 49.

Dependent on the number of completed trials, some teachers ended up with different numbers of overlaps (e.g. overlap between round 1 and 2 for teacher P1 because there was no trial in the third round of implementation). For further analysis at the level of the individual teacher, this posed the problem that there was not the same number of data points for every teacher (e.g. data points from one overlap for teacher P1 but data point from two overlaps for teacher P3). Comparing the overlap between rounds 1 and 2 with the overlap between rounds 2 and 3 within teachers, the differences in the sum are small (below 1 in almost all cases). For further analysis of the data, the overlaps between the first two subsequent rounds per teacher were therefore considered only (e.g. only the overlap between round 1 and round 2 for teacher P3). The resulting distribution can be found in Table 50.

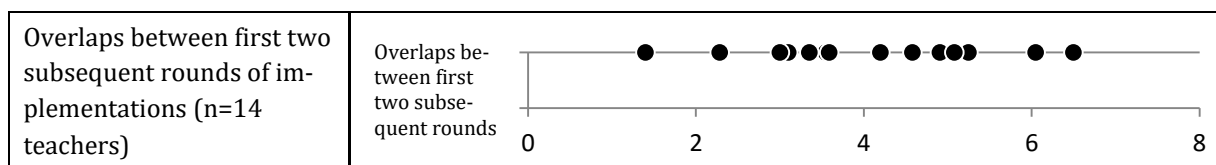
Table 49: Variability of implementations within teachers.

O1 = Overlap of trials in dimensions of openness [%]
 O2 = Overlap of trials in inquiry activities [%]
 O3 = Overlap of trials in competences assessed [%]
 O4 = Overlap of trials in communication of criteria [%]
 O5 = Overlap of trials in sources of data [%]
 O6 = Overlap of trials in assessment methods [%]
 O7 = Overlap of trials in engagement with feedback [%]
 O8 = Overlap of trials in cycle length [%]
 Sum = sum of O1, O2, O3, O4, O5, O6, O7, and O8

	Overlap between rounds 1 and 2									Overlap between rounds 2 and 3									
	O1	O2	O3	O4	O5	O6	O7	O8	Sum	O1	O2	O3	O4	O5	O6	O7	O8	Sum	
Primary school teachers	P1	0.75	0.43	0.40	0.00	0.50	0.00	0.50	1.00	3.58	--								
	P2	0.25	0.43	0.40	1.00	0.50	0.50	1.00	1.00	5.08	--								
	P3	0.40	0.43	0.25	1.00	0.33	0.50	1.00	1.00	4.91	0.75	0.75	0.25	1.00	0.00	0.00	1.00	1.00	4.75
	P4	--									--								
	P5	0.50	0.25	0.50	1.00	0.00	1.00	1.00	1.00	5.25	--								
	P6	0.00	0.40	0.00	0.00	0.00	0.00	0.00	1.00	1.40	--								
	P7	--									--								
	P8	0.50	0.50	0.20	1.00	0.50	0.00	0.50	1.00	4.20	--								
	P9	0.50	0.60	0.00	0.00	1.00	0.00	0.00	1.00	3.10	0.33	0.80	0.50	0.00	1.00	1.00	0.00	0.00	3.63
Upper secondary school teachers	S1	1.00	0.80	0.25	1.00	1.00	0.00	1.00	1.00	6.05	0.50	0.83	0.25	1.00	1.00	0.00	1.00	1.00	5.58
	S2	0.60	0.40	0.29	0.00	0.00	0.50	0.50	0.00	2.29	0.60	0.29	0.25	0.00	0.00	1.00	0.50	0.00	2.64
	S3	--									1.00	0.25	0.33	1.00	0.00	0.00	1.00	0.00	3.58
	S4	1.00	0.50	1.00	1.00	1.00	0.00	1.00	1.00	6.50	1.00	0.67	0.50	1.00	1.00	1.00	1.00	1.00	7.16
	S5	0.50	0.75	0.33	1.00	1.00	0.00	0.50	0.50	4.58	0.67	0.71	0.25	1.00	1.00	0.50	1.00	1.00	6.12
	S6	--									--								
	S7	0.40	0.28	0.17	1.00	0.50	0.00	1.00	0.00	3.35	0.50	0.60	0.25	1.00	0.00	0.00	0.00	0.00	2.35
	S8	--									0.00	0.00	0.00	1.00	1.00	0.00	1.00	0.00	3.00
	S9	--									--								
	S10	--									--								
	S11	--									--								

The graphic representation illustrates that the overlaps of the teachers vary extensively. The data available covers almost the whole spectrum that would be theoretically possible (theoretical minimum: 0; theoretical maximum: 8). From the data in Table 49 and from its graphical representation in Table 50 it is, however, not clear, whether this distribution of overlaps follows a particular pattern.

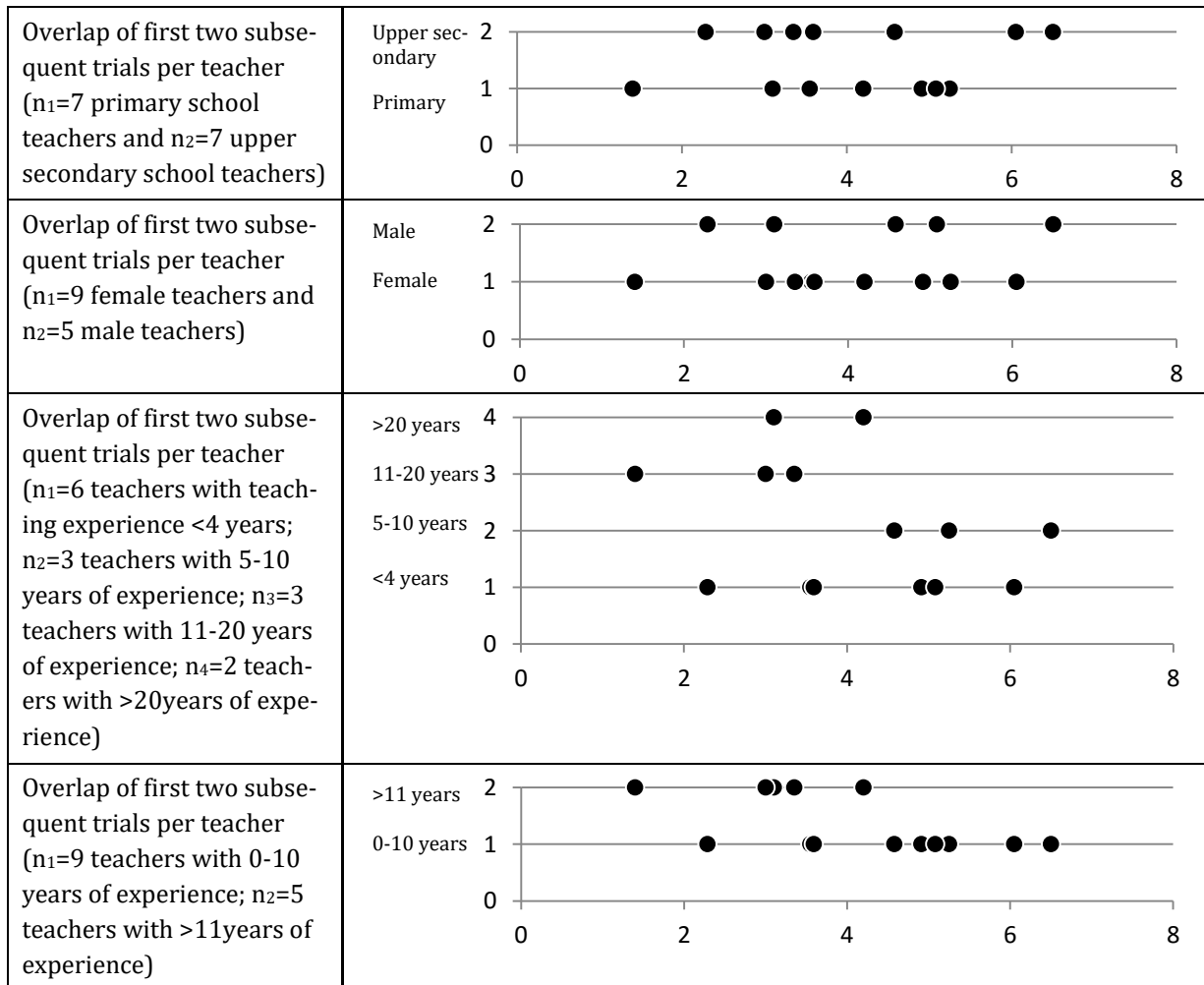
Table 50: Overlaps



In order to explore possible patterns that might partially explain the implementation habits of the teachers in the study, the data from Table 49 were grouped according to different variables: The school level; the teachers' gender; and the teachers' teaching experience (see Table 5 and Table 6). A first visual impression of the results is provided in Table 51. These representations show that all variables explored might poten-

tially have an influence on the implementation habits of the teachers. For the teaching experience, the pattern with four groups (n=6 teachers with teaching experience <4 years; n=3 teachers with 5-10 years of experience; n=3 teachers with 11-20 years of experience; n=2 teachers with >20years of experience) was not entirely clear. Therefore, the four groups were also collapsed into two groups only (n=9 teachers with 0-10 years of experience; n=5 teachers with >11years of experience).

Table 51: Overlap of first two subsequent trials per teacher grouped by different variables.



Since all variables explored in Table 51 appeared to potentially be part of an explanation for the teachers' different implementation habits, Mann-Whitney-U-Tests were performed. The results can be found in

Table 52 (for the two school levels); and

Table 53 (for the teachers' gender);

Table 54 (for the teachers' teaching experience).

The following

Table 52 will show the results of the tests on significant differences in the implementation habits between the teachers from the two school levels.

Table 52: Descriptive statistics and significance tests of the overlaps, grouped by school level. Mdn=median; AM= arithmetic mean; SD= standard deviation. Differences between the trials from the two school levels tested by Mann U Whitney (*p<0.05; **p<0.01) and effect sizes with Cohen's d; |0.2|<d≤|0.5| represents small effect size; |0.5|<d≤|0.8| represents medium effect size; |0.8|<d represents large effect size (Cohen, 1988).

	Primary school teachers (n ₁ =7)			Upper secondary school teachers (n ₂ =7 teachers)			Mann-Whitney-U		Effect size
	Mdn	AM	SD	Mdn	AM	SD	U	p (2*(1 tailed))	d
O1 = Overlap of trials in dimensions of openness	0.50	0.43	0.22	0.60	0.65	0.32	13.00	n.s.	0.80
O2 = Overlap of trials in inquiry activities	0.43	0.47	0.14	0.43	0.51	0.26	23.50	n.s.	0.19
O3 = Overlap of trials in competences assessed	0.25	0.29	0.15	0.27	0.32	0.21	24.50	n.s.	0.16
O4 = Overlap of trials in communication of criteria	1.00	0.57	0.50	1.00	0.83	0.35	17.50	n.s.	0.60
O5 = Overlap of trials in sources of data	0.50	0.38	0.33	1.00	0.63	0.46	18.00	n.s.	0.62
O6 = Overlap of trials in assessment methods	0.25	0.32	0.35	0.00	0.25	0.28	20.50	n.s.	-0.22
O7 = Overlap of trials in engagement with feedback	0.50	0.57	0.42	1.00	0.79	0.22	17.00	n.s.	0.66
O8 = Overlap of trials in cycle length	1.00	0.92	0.17	0.00	0.46	0.46	9.00	n.s.	-1.33
Sum of O1 to O8	4.20	3.96	1.07	3.58	4.27	1.54	22.50	n.s.	0.23

The results from

Table 52 show that with Mann Whitney- U tests, no statistically significant changes between the primary- and the upper secondary school teachers participating in the study were found. Looking at Cohen's d, effects were found for almost all variables: Medium-sized effects in the overlap in dimensions of openness, in communication of criteria, in sources of data, and in the engagement with feedback; a small effect in the assessment methods; and a large effect in the cycle length. In the sum of all types of overlap, a small effect of the school level on the implementation habits was found.

In the next part of the analysis, the potential influence of the teachers' gender on the implementation behaviour will be explored. The following

Table 53 will show the results of the tests on significant differences in the implementation habits between the teachers' gender.

Table 53: Descriptive statistics and significance tests of the overlaps, grouped by teachers' gender. Mdn=Median; AM= arithmetic mean; SD= standard deviation. Differences between the trials from the two school levels tested by Mann U Whitney (*p<0.05; **p<0.01) and effect sizes with Cohen's d; |0.2|<d≤|0.5| represents small effect size; |0.5|<d≤|0.8| represents medium effect size; |0.8|<d represents large effect size (Cohen, 1988).

	Female teachers (n ₁ =9)			Male teachers (n ₂ =5)			Mann-Whitney-U		Effect size
	Mdn	AM	SD	Mdn	AM	SD	U	p (2*(1 tailed))	d
O1 = Overlap of trials in dimensions of openness	0.50	0.50	0.33	0.59	0.57	0.28	20.50	n.s.	0.21
O2 = Overlap of trials in inquiry activities	0.43	0.41	0.23	0.59	0.56	0.17	14.00	n.s.	0.71
O3 = Overlap of trials in competences assessed	0.25	0.24	0.17	0.29	0.39	0.21	11.50	n.s.	0.85
O4 = Overlap of trials in communication of criteria	1.00	0.78	0.44	1.00	0.60	0.55	18.50	n.s.	-0.37
O5 = Overlap of trials in sources of data	0.25	0.38	0.40	1.00	0.70	0.45	13.50	n.s.	0.77
O6 = Overlap of trials in assessment methods	0.00	0.14	0.33	0.50	0.50	0.18	5.50	p<0.05	1.24
O7 = Overlap of trials in engagement with feedback	1.00	0.72	0.36	0.75	0.65	0.42	20.00	n.s.	-0.19
O8 = Overlap of trials in cycle length	1.00	0.67	0.50	0.75	0.65	0.42	19.50	n.s.	-0.04
Sum of O1 to O8	3.58	3.83	1.36	5.08	4.61	1.71	17.00	n.s.	0.52

The results from

Table 53 show that with Mann Whitney- U tests, statistically significant differences between male and female teachers could be found in the overlap of trials in assessment methods (O6). No significant differences could be found in the other overlaps. Medium- and large Cohen effect sizes were found in O2, O3, O5, O6, and in the sum of O1 to O8.

The female teachers in the study had a smaller overlap between trials in terms of the assessment methods (median 0.00; low values stand for low overlap) than male teachers (median 0.50; exact Mann-Whitney-U test: $U(n_1=9, n_2=5)=5.50; p<0.05$). Cohen's effect size was $d=1.24$ which represents a large effect.

In the next step of analysis, the potential influence of the teaching experience on the implementation behaviour was explored. Because of the small size of the four sub-samples defined by the teaching experience (see Table 51), the four sub-samples were collapsed into two sub-samples: Teachers with a teaching experience between 0-10 years and teachers with a teaching experience above 10 years. The results on the respective analysis can be found in

Table 54 below.

Table 54: Descriptive statistics and significance tests of the overlaps, grouped by teaching experience. Mdn=median; AM= arithmetic mean; SD= standard deviation. Differences between the overlaps of the two teacher groups tested by Mann U Whitney (*p<0.05; **p<0.01) and effect sizes with Cohen's d; |0.2|<d≤|0.5| represents small effect size; |0.5|<d≤|0.8| represents medium effect size; |0.8|<d represents large effect size (Cohen, 1988).

	Teaching experience 0-10 years (n ₁ =9)			Teaching experience >11 years (n ₂ =5)			Mann-Whitney-U		Effect size
	Mdn	AM	SD	Mdn	AM	SD	U	p (2*(1 tailed))	d
O1 = Overlap of trials in dimensions of openness	0.60	0.67	0.24	0.41	0.27	0.25	3.50	p<0.01	-1.64
O2 = Overlap of trials in inquiry activities	0.43	0.49	0.20	0.43	0.41	0.26	19.00	n.s.	-0.40
O3 = Overlap of trials in competences assessed	0.33	0.38	0.16	0.20	0.13	0.12	1.00	p<0.01	-1.67
O4 = Overlap of trials in communication of criteria	1.00	0.78	0.44	1.00	0.60	0.55	18.50	n.s.	-0.37
O5 = Overlap of trials in sources of data	0.50	0.46	0.45	0.50	0.55	0.45	19.50	n.s.	0.20
O6 = Overlap of trials in assessment methods	0.25	0.36	0.36	0.00	0.10	0.22	12.00	n.s.	-0.82
O7 = Overlap of trials in engagement with feedback	1.00	0.86	0.22	0.50	0.40	0.42	8.00	n.s.	-1.53
O8 = Overlap of trials in cycle length	1.00	0.75	0.43	0.50	0.50	0.50	16.00	n.s.	-0.55
Sum of O1 to O8	5.08	4.75	1.33	3.00	2.96	1.02	6.00	p<0.05	-1.45

The results from

Table 54 show that with Mann Whitney- U tests, statistically significant differences between the implementation habits of the teachers with less teaching experience and those with more teaching experience in the study could be found in the overlap of trials in dimensions of openness (O1), in the competences assessed (O3), and in the sum of all overlaps (sum of O1 to O8). No significant differences could be found in the other overlaps. As Table 51 shows, these results depend on the formation of the subgroups: If, for example, the less experienced teachers were defined as 0-4 years of teaching experience and the more experienced teachers as >4 years of teaching experience, the differences disappeared. Medium- and large effect sizes were found in O1, O3, O6, O7, O8, and in the sum of O1 to O8.

The less experienced teachers in the study had a higher overlap between trials in terms of the dimensions of openness (median 0.60; low values stand for low overlap) than more experienced teachers (median 0.41; exact Mann-Whitney-U test: $U(n_1=9, n_2=5)=3.50$; $p<0.01$). Cohen's effect size was $d= -1.64$ which represents a large effect.

The less experienced teachers in the study also had a higher overlap between trials in terms of the competences assessed (median 0.33; low values stand for low overlap) than more experienced teachers (median 0.20; exact Mann-Whitney-U test: $U(n_1=9, n_2=5)=1.00$; $p<0.01$). Cohen's effect size was $d= -1.67$ which represents a large effect.

Finally, the less experienced teachers in the study had a higher overall overlap between trials (median 5.08; low values stand for low overlap) than more experienced teachers (median 3.00; exact Mann-Whitney-U test: $U(n_1=9, n_2=5)=6.00$; $p<0.05$). Cohen's effect size (1988) was $d=-1.45$ which represents a large effect.

In their implementation habits, differences between different teachers in the study were found: Whereas some teachers had high variabilities between their trials, others had a high overlap between one trial and the subsequent one. The extent of the variability remained stable for the individual teachers across the three rounds of implementation. Despite intermediate and large effect sizes, no significant differences were found between the teachers of the different school levels. At the level of gender, significant differences were found in the variability of assessment methods: The female teachers collaborating in the study varied the assessment methods trialled significantly more than the male teachers. The respective effect size was large. Significant differences were also found between the teachers with less than ten years of teaching experience and the teachers with a higher teaching experience in the study: The more experienced teachers varied their inquiry contexts significantly more than their less experienced peers. The effect sizes were large again. The last results, however, depend on the formation of the subgroups: If, for example, the less experienced teachers were defined as 0-4 years of teaching experience and the more experienced teachers as >4 years of teaching experience, the differences disappeared.

8 Discussion

The discussion will be structured as follows: In sub-chapters 8.1 – 8.4, the data relating to the four research questions will be summarized, discussed and compared to the research literature as introduced in chapter 3. In the following, sub-chapters 8.5 and 8.6 will focus on overarching themes: Measures of support for formative assessment practices in Switzerland (sub-chapter 8.5), and teachers developing their own formative assessment practices in the context of this study (sub-chapter 8.6).

This is an exploratory study that aims at generating hypotheses on the implementation of formal formative assessment methods in inquiry-based science education in Switzerland. The respective hypotheses will be deduced in sub-chapters 8.5 and 8.6.

8.1 Discussion of research question 1: The teachers' understanding of formative assessment

In this sub-chapter, the results of research question 1 '*What is the teachers' understanding of formative assessment*' will be summarized, discussed and compared to the literature.

Summary of results (copied from the last part of 7.1)

The teachers' descriptions of the term 'formative assessment' were analysed. The teachers were asked to write down an explanation of the expression three times during the study (beginning of collaboration; middle phase; end of collaboration). From all the answers, 11 categories were developed: Five categories that describe elements also present in the literature on formative assessment (supportive nature; guidance about next steps in learning or in teaching; individual; prospective nature); one element that is generally ascribed to assessment in the literature (criterion-based nature); four categories with other elements (focused on a specific set of competences such as behaviour in groups; individual reference norm; grading of the learning process; unclear or reference to inquiry) and one category with examples of assessment methods.

Implications

The results show that the majority of the teachers from both school levels who collaborated in the study described the term 'formative assessment' similar to what a definition in the literature on formative assessment would look like – from the beginning of the study. The high overlap could be interpreted as a sign of congruence between research and practice. A small portion of the teachers' descriptions confused formative assessment with other concepts such as assessment of so-called soft skills. Two misconceptions that are rather common in Switzerland were present in the teachers' descriptions from both school levels as well: That formative assessment is related to grading of the learning process (rather than the product) and that formative assessment is individually referenced. Finally, the references to inquiry and to the assessment methods could origin from the setting of the study which took place in the context of inquiry-based education and involved the trial of three examples of assessment methods.

The opinion that formative assessment and respective feedback could cover criticism exclusively or the opposite; that it could only contain praise, was not found in the results.

With respect to support measures for an implementation of formative assessment methods (see sub-chapter 8.5), the two misconceptions found in this study should be addressed: That formative assessment is related to grading of the learning process rather than the product and that formative assessment is individually referenced.

Comparison to the literature

It is not easy to compare the results on the teachers' descriptions of 'formative assessment' from this study to the literature: Teacher concepts of formative assessment have been poorly investigated. A paper on teacher concepts of assessment in general (not particularly on formative assessment) was found (Brown, 2004). It reveals that teachers see three different purposes of student assessment which are congruent to

the purposes mentioned in the literature: Improving the students' learning in the classroom (formative assessment); checking the students' learning (summative assessment); control the quality of a teacher or a school (evaluation). A second paper (Brown et al., 2012) focusses on teachers' concepts of feedback and its relation with formative assessment. The main outcome is that the teachers investigated considered the feedback as a part of formative assessment, by guiding the students' learning, rather than only as a means to offer praise. So the detailed results of the two papers cannot be easily linked to this study. However, two general outcomes can be related to this study: The teachers' conceptions of assessment and of the purpose of feedback could be well-described by elements and categories from the research literature. Similar results were also found in this study. This can be taken as a sign for the high research-practice congruence in the understanding of assessment. The second general outcome was found in Brown et al., 2012. The authors find only minor differences between the teacher concepts at different school levels. The same result was found in this study. This can be taken as a sign that teacher concepts, in broad, may depend on other variables than the school level.

8.2 Discussion of research question 2: Description and analysis of the teachers' trials

In this sub-chapter, the results of research question 2 '*How do the teachers in the study trial formative assessment methods in their inquiry-based science education?*' will be summarized, discussed and compared to the literature.

8.2.1 Inquiry units in the trials

Summary of results (copied from the last part of 7.2.1)

The units trialled were characterized by three criteria: The dimension(s) of openness; the inquiry activities enacted in the units; and the competences assessed.

Openness refers to the idea that in inquiry-based education, not all aspects are pre-defined but some decisions are left to the students. These decisions may concern different dimensions. The dimensions of openness were conceptualized after Priemer (2011; see sub-chapter 3.1 for details). The analysis of the trials in this study shows that all dimensions of openness were covered by at least a few trials at both school levels. Whereas only few inquiry units were open in terms of 'content' and 'strategies', more inquiry units were open in terms of 'methods'. Almost all inquiries were open in terms of 'solution' and 'solution process'. The differences between school levels were small. Many trials covered more than one dimension of openness: The peak at primary school was around units which were open in three dimensions. The distribution at upper secondary school had a maximum at two open dimensions and another maximum at open in five dimensions.

Looking at the inquiry activities enacted in the units, it appears that all activities as defined in Bell et al. (2010) were part of at least one trial in both school levels. However, huge differences in frequency occur: Whereas 'orienting and asking questions'; 'hypothesis generation'; 'model'; and 'prediction' were rarely part of the inquiries at both school levels, 'planning'; 'investigation'; 'analysis and interpretation'; and 'communication' were frequently enacted at both school levels. 'Conclusion and evaluation' was often part of the inquiries at upper secondary school but not at primary school. Looking at the number of inquiry activities per unit, the peak of the primary school units is around four inquiry activities. At upper secondary school, there is no clear peak, but most trials included between 4 and 6 activities.

Both domain-specific and transversal competences are ascribed to inquiry-based education (see sub-chapter 3.1 for details). In this study, the conceptualisation of domain-specific competences that are fostered by inquiry from Bell et al. (2010) were taken as a basis. The results show that all domain-specific competences were assessed at least once in the trials. However, there are differences in the frequency of occurrence: At primary school, 'orienting and asking questions', 'model', 'conclusion and evaluation', as well as 'prediction' was not assessed at all. 'Hypothesis generation', 'planning', and 'analysis and interpretation' were rarely assessed. By far the most-assessed competences in the primary school trials were 'investigation' and 'communication'. At upper secondary school, 'hypothesis generation' was not assessed but all other competences were. 'Orienting and asking questions', 'model', 'conclusion and evaluation', and 'prediction' were rarely assessed. 'Planning', 'investigation', and 'analysis and interpretation' appeared at a moderate frequency. By far the most-assessed competence was 'communication'. The results also show that trials with one domain-specific competence assessed were most frequent at both school levels. Two, three or four competences occurred less frequently. At upper secondary school, there was a small number of trials with 5 or 6 competences assessed.

When deciding about what domain-specific competences to assess, the decision-making process of the teachers seems to take place on different levels: At primary school level, the most frequently mentioned line of argumentation included no explicit decision. Instead, the teachers explained that the competences for assessment emerged naturally during the preparation of the unit. Less frequently, the teachers brought up resource-based decisions. Finally, some teachers chose a particular competence because they thought it was important for science education. At upper secondary school, the two most commonly mentioned lines of

argumentations were that a particular competence was considered important for the students' further career or generally important in science education. Less frequently, the decision was taken based on the students' abilities.

The analysis of the transversal competences assessed was based on the conceptualizations from OECD (2005b, see sub-chapter 3.1 for details). The results show that transversal competences were assessed in most trials at primary school and in half of the trials at upper secondary school. They were always assessed in combination with at least one domain-specific competence. The teachers' reasons for deciding on a particular transversal competence were driven by the perceived relevance of this competence at both school levels.

Implications and attempts to explain the results with traditions in Swiss classrooms

Priemer (2011) arranged the dimensions of openness in a hierarchical order (with openness in the content being the highest-order dimension and openness with respect to the solution process being the lowest-order dimension). In the trials in this study, the frequency of coverage ascends with decreasing order of dimension (e.g. openness with respect to content and strategies is rare; openness with respect to the number of solutions and to the solution processes is frequent). This could be a sign that not all dimensions of openness are equally easy to implement in practical units. Openness in terms of content and strategy at upper secondary schools, for example, seems to be possible in *Maturaarbeiten* (thesis at the end of upper secondary school) almost exclusively whereas at primary school level, students can be given the choice between different animals, parts of the body etc. to focus on. Openness in terms of solutions or solution processes on the other hand seems to be more common in inquiry units at both school levels. Quantitative data from the compulsory school levels (HarmoS, 2008) suggests, however, that a large portion of primary school teachers do not often work in open settings such as inquiry teaching.

Similarly, not all inquiry activities as defined in Bell et al. (2010) were enacted in the same frequency. Again, it could be a sign that some activities appear easier to be realised to teachers at the two school levels explored: Activities like 'orienting and asking questions'; 'hypothesis generation'; 'model'; and 'prediction' make most sense in an extensive inquiry unit where students deeply immerse into a particular topic. Other activities such as 'investigation' and 'documentation' of respective data could easily take place in a shorter unit, where the main aim lies in exemplifying a particular law to complement its theoretical introduction. A study on the use of experiments in physics at lower secondary school in Switzerland (Börlin, 2010) provides further evidence for this explanation: He found that many experiments take place in a short time, and that the conduction is central whereas little weight is given to the research questions, respective hypotheses, and to reflections on results. An underlying mechanism to explain the results from this study could therefore be the role of experiments in science education: If their aim often lies in the exemplification and illustration of theoretical explanations, this could also affect the focus of inquiry-based education. Part of this affection could be that the conduction and documentation are considered more relevant to be practiced than other parts of an investigation or experiment. The difference in the number of inquiry activities enacted in a particular unit between the two school levels could be caused by the difference in the working speed of the respective students.

Moving from the inquiry activities enacted to the domain-specific competences assessed, the priorities seem to narrow. At primary school level, the investigation competence was assessed most frequently. This gives the impression that at this school level, a competence that can be diagnosed by direct observation (rather than diagnosis based on a written report or similar) is most feasible for assessment. It could also be that a practical competence such as investigation is easier to operationalise than a competence like hypothesis generation where it might be more difficult to elaborate concrete indicators for diagnosis. A third aspect to explain the results at primary school level could be that hands-on activities appeared more relevant than others and are therefore given more weight in assessment. At upper secondary school, the competence assessed most frequently was communication which also included the documentation of inquiries in lab reports; and similar activities. Since this is an activity typically assessed for summative purposes, it might be obvious to involve it in formative assessment as well.

Many of the teachers did not focus on only one specific aspect but rather aimed at a certain broadness of criteria in their assessments. This could be, again, an influence from summative assessment: Non-traditional summative assessment (i.e. not classic paper-and-pencil-tests but project presentations and similar) are likely to cover a broad range of criteria (Widmer Märki, 2011). The cases at upper secondary school with rather holistic assessments (where five or six competences were involved) took place in the context of *Maturaarbeiten*. Trying to interpret the teachers' reasoning for the competences chosen for assessment, the teachers from both school levels seemed to choose the domain-specific competences based on their personal conviction on what competences are important or, at primary school level, without a conscious decision-making process. Considerations on the concrete inquiry unit or references to the curriculum were not mentioned.

Many of the teachers involved transversal competences in their formative assessment, particularly at the primary school level. This is remarkable because they were not explicitly told to do so for the purposes of the study. Furthermore, transversal competences are not typically part of the summative assessment practices at any school level in Switzerland. Nevertheless, all three categories of transversal competences (to use tool interactively; to interact in heterogeneous groups; to act autonomously; OECD, 2005b) were assessed in several trials at both school levels. The teachers' reasoning at both school levels revealed that particular transversal competences were chosen based on the teachers' personal conviction on what transversal competences are important but are not assessed in classical tests. The teachers' reasoning for both domain-specific and transversal competences can be considered a sign of the high autonomy they have in their teaching, as described in sub-chapter 3.8.

Comparison of the results with curricula

In this section, the inquiry activities enacted and assessed in the study will be compared to the guidelines in the curricula from compulsory school level and *Gymnasium* as introduced in sub-chapter 3.8. Of course such a comparison is difficult since the results of the study only provide an insight into one exemplary unit per semester and therefore cannot give a real overview. Furthermore, curriculum 21 for the compulsory school levels is not yet valid for all *Kantone* yet (in 2016). The interpretations will therefore be tentative.

As shown in sub-chapter 3.8, both the competence model for compulsory school science education (HarmoS, 2008) and the curriculum 21 (D-EDK, 2014) cover the inquiry activities as defined in Bell et al. (2010), apart from 'prediction' which is not part of the national documents. The impression from the activities enacted in the primary school trials is that some of the aspects within the skill 'to ask questions and to investigate' from the competence model are trained much more frequently than others: Whereas 'to pose questions, problems and hypotheses' and 'to reflect upon results and methods' were rarely part of the inquiries, 'to choose and use suitable tools, instruments and materials' as well as 'to conduct explorations, investigations or experiments' were trained in many inquiries. The aspect of the skill 'to communicate and to exchange' which appeared to be closely associated with inquiry-based education as defined in Bell et al. (2010), 'to describe, present and reason' was also frequently included in the units trialled. Looking at the competences which were not only trained but also formatively assessed in the trials of the study, the emerging picture is that 'to conduct explorations, investigations or experiments' from the skill 'to ask questions and to investigate' and 'to describe, present and reason' from the skill 'to communicate and to exchange' are much more frequently assessed formatively than the other aspects of the skill 'to ask questions and to investigate'. With respect to the implementation of the curriculum 21 in all *Kantone*, this uneven distribution in the teaching practice as reflected in the trials of this study should be considered.

At upper secondary school level, the picture is more difficult to interpret since the skills are spelled out separately for every subject (physics, chemistry, biology; see Table 3 in sub-chapter 3.8) in the curriculum whereas the results of this study were not analysed per subject. Furthermore, the skills in the curriculum are formulated in a more abstract way than the competences in Bell et al. (2010). What can be said is that generally, some of the skills formulated in the curricula for the three science subjects at *Gymnasium* level appear to be much more trained and formatively assessed than others. Examples of the frequently trained

and assessed skills include 'to plan and conduct meaningful experiments [...], to record and represent data in words and graphically, [...]' in the biology skills (EDK, 1994; see sub-chapter 3.8). The rarely trained and assessed skills include 'to develop models and apply them on specific situations' in the physics skills (EDK, 1994; see sub-chapter 3.8).

8.2.2 Formative assessment in the trials

Summary of results (copied from the last part of 7.2.2)

The teachers' formative assessment activities were characterized in terms of different aspects: The communication of the criteria, the data sources used for diagnosis, the assessment methods, the means of engaging with the feedback, and the length of the assessment cycles.

As part of the formative assessment, the assessment criteria were introduced in the trials. Three ways of introduction were found in the study: Most frequently at both school levels, the assessment criteria were pre-defined and explicitly communicated by the teacher. At primary school level, the assessment criteria were elaborated with the students in two cases. In a few cases at both school levels, the criteria were implicitly clear (for example in some of the cases where documenting experimental results in the lab journal are regular part of the lab lessons).

For diagnosis, a number of types of data on student learning were used. Amongst those, observational data was most frequent in the primary school trials whereas written student data was most common in the upper secondary school trials. In the primary school trials, one or two sources of data were used whereas in the upper secondary school trials, the use of only one source of data was most common.

Three formal formative assessment methods were trialled (written teacher assessment; peer-assessment; self-assessment) several times at both school levels. At both school levels but more often at primary school level, more than one assessment method was embedded in one trial even though that was not part of the teachers' task in the study. The relation between assessment methods and competences assessed was analyzed, but no clear pattern emerged: All three assessment methods were used to assess both domain-specific and transversal competences.

In the individual interviews, the teachers were asked how they chose a particular assessment method for their trials. A number of decision-making processes could be revealed: At both school levels, some teachers answered that there was no particular reason for their choice but that the formative assessment just appeared suitable in the context of a particular situation. A second line of argumentation at both school levels was related to the teachers' confidence to work with a particular method. At primary school level, the third reason was the students' motivation to work with a particular method. At upper secondary school level, the learning benefits of using the method itself (for example reflective skills that develop from self-assessment) and organizational issues were also mentioned.

As a next step, the means of engaging with the feedback received was analyzed. The two options mentioned in the literature (see sub-chapter 3.2) are revision of the original artefact/activity or transfer to a subsequent artefact/activity. The results from the study showed that at primary school, the feedback was more likely to be used for revision whereas at upper secondary school, the feedback rather enhanced the subsequent work (e.g. the next lab report). A second effect that was visible was that feedback on domain-specific competences was more often used for revision whereas feedback on transversal competences was more likely to be transferred to subsequent units.

The cycle length of the formative assessment activities trialled (see sub-chapter 3.2) was analyzed. The results showed that in the study, the primary school teachers' cycles were typically shorter (minute by minute, day by day) whereas the upper secondary school teachers implemented longer cycles (1-4 weeks or 4 weeks to 1 year). A second effect that was visible in the study was that feedback on domain-specific competences

was typically used in shorter cycles whereas feedback on transversal competences was more likely to be used in longer cycles.

Implications and attempts to explain results based on traditions in Swiss classrooms

A possible explanation for the pre-defined assessment criteria is that the teacher defining and communicating the criteria is the most rapid way to get the formative assessment process started whereas elaborating criteria with students might take more lesson time. It could be that the teachers in the study did not consider the introduction of the assessment criteria the most relevant part of the formative assessment and that they therefore decided to rather dedicate the lesson time to diagnosis and feedback. A third possible explanation is that the teachers were simply not aware of the possibility to have students elaborate assessment criteria and indicators to judge their pieces of work themselves.

The data used for diagnosis seems to correlate with the competences assessed (see 7.2.1 and 8.2.1): At primary school, the choice of observational data seems appropriate to assess the hands-in investigative part of an inquiry (see section on competences assessed). The choice of often two sources of data may reflect the holistic approach being more common amongst the primary school teachers with their pedagogical socialization (compared to the subject-orientation of the upper secondary school teachers in the study; see subchapter 5.2). At upper secondary school, the use of written student data to assess the documentation of an investigation or an experiment could be obvious for the teachers because they will also use them for summative assessment: Summative assessment will typically need strong, undeniable evidence such as written student reports rather than observations which cannot be reconstructed later on. This could reflect an approach focused on grades. A study on summative assessment in interdisciplinary science units at upper secondary school from Widmer Märki (2011) also found that the teachers used mostly written student data for their assessment.

The fact that all three formative assessment methods were trialled by teachers from both school levels illustrates, together with the variety of enactments, the appropriateness of these formative assessment methods for the Swiss educational context at the two different school levels and in particular in the context of inquiry. Some of the trials involved more than one assessment method. All teachers who worked with more than one assessment method in the same trial used these assessment methods to assess different competences. This same result was also found in the above-mentioned study on summative assessment from Widmer Märki (2012). It implies that these teachers had a certain expectation of what assessment method would fit which competence (e.g. what competence could be assessed by peers or what competence could be reflected on).

From the data of the study, there is a slight tendency that written teacher assessment and peer-assessment were often used to assess domain-specific competences whereas self-assessment was often used to assess transversal competences. This seems plausible because the transversal competences as defined in the study might be more difficult to diagnose by other people than by the assessor him/herself.

Overall, the decision-making process of the teachers on what assessment method to choose seems to be situated on two levels: In some cases, the choice of an assessment method was not a conscious decision but part of teaching instinct. In other cases, the specific situation was taken into consideration: The unit; the abilities and the motivation of a particular class; the self-efficacy of the teacher; and practical issues.

There were considerable differences in the use of the feedback between the two school levels. This seems plausible because the transfer itself might be a hard task for a primary school student but might be more appropriate for the abilities of upper secondary school students. Another factor could be that the time pressure at primary school is lower so that students can be given the time to revise a piece of work. Furthermore, the recurrent lab lessons and the associated lab reports at upper secondary school were often used for the trials at upper secondary school. The formative assessment of one lab report and the use of the respective feedback in the subsequent lab reports appears obvious under these circumstances. Looking at the use of feedback in more detail, the results showed that the revision of original artefacts/activities usually referred

to domain-specific competences whereas the transfer to subsequent artefacts/activities typically concerned transversal competences. This seems plausible too, since the transversal competences almost exclusively related to a particular situation (behavior in a group; autonomy in the context of a particular task) rather than an artefact and therefore could not be revised.

Differences between the two school levels were also found in the cycle length of the formative assessment. One part of the explanation of these patterns is certainly the differences in the lesson structures. At primary school, the students typically have four single science lessons or two double science lessons per week whereas at upper secondary school, the lab lessons (where many of the trials for the study took place) typically recur every second week only. Another part of the answer is revealed in the above section on the use of the feedback: At upper secondary school, the use of feedback often takes place in the form of a transfer to a subsequent activity rather than in the revision of an ongoing activity (which was more common in the primary school trials for the study) which will potentially result in longer cycles.

8.2.3 Problems in the trials

Summary of results (copied from the last part of 7.2.3)

Nineteen out of 53 cases collected in the study did not match the criteria as specified in chapter 5.4. These criteria were: (1) Conduction of trial; (2) sufficient documentation of trial for analysis; (3) trial in the context of inquiry-based education; (4) trial involving formative assessment.

Seven teachers from both school levels did not manage to trial anything even though they were paid for it and motivated to do it. They all said that they did not find time during the semester to conduct a formal formative assessment activity in the context of inquiry due to the extensive curriculum and due to the many other activities (such as teachers' military service or student exchange programs).

Three primary school teachers' trials were not documented in a way they could be analyzed which is a methods issue.

In the case of a teacher from primary school and a teacher from upper secondary school, the problem was not with formative assessment but with inquiry. The teacher from upper secondary school was aware of the fact that she was not following the instructions given for the study but the teacher at primary school was not.

A few teachers from both school levels did not trial formative assessment activities. One of the teachers simply forgot the formative assessment part because he was so busy with the inquiry. The other teachers either did not clarify the assessment criteria at the beginning of the formative assessment activity or they did not provide the students with the opportunity to make use of the feedback they received. Some of them were not aware of the issues with their trials whereas others mentioned them during the group discussions.

Implications and comparison to the literature

The results show that time seems to be a substantial barrier for formal formative assessment, even for motivated teachers. This is consistent with references in the literature (see sub-chapter 3.6; OECD, 2005a; Looney, 2011). The problem will be discussed in more detail in sub-chapter 8.5 on potential measures of support for teachers.

The difficulties related to the inquiry setting show, firstly, that the teachers had different concepts of what the term 'inquiry' means. It also shows another methods issue: The teachers' task in the study, to trial a formative assessment method in their inquiry-based education, really consisted of two challenges (firstly inquiry and secondly formative assessment). This will be discussed in more detail in sub-chapter 9.2, the critique of methodology.

The problems that occurred with formative assessment show that the teachers in the study were aware of the importance of diagnosis and provision of feedback but, in some, cases, forgot the other two parts of formative assessment. The finding sharpens earlier claims (see sub-chapter 3.6 on the abilities of the teachers; e.g. Bennett, 2011; Cizek, 2010; Stiggins, 1999) where insufficient assessment literacy of the teachers was reported. The problem will be discussed in more detail in sub-chapter 8.5 on potential measures of support for teachers.

8.3 Discussion of research question 3: Teachers' and students' evaluations of the methods trialled

In this sub-chapter, the results of research question 3 '*How do the teachers and the students evaluate the formative assessment methods trialled?*' will be summarized, discussed and compared to the literature.

8.3.1 Usability of the methods for different school levels as perceived by the teachers

Summary of the results (copied from the last part of 7.3.1)

All three formative assessment methods trialled in the study were considered usable at their respective school levels by the teachers in the study. Reasons for these evaluations can be found in the subsequent sections.

Implications and comparison to the literature

This can be taken as a sign that the assessment methods were generally accepted as a valuable part of teaching by the teachers of both school levels. The results can be related to earlier findings which confirm a generally positive teacher attitude towards formative assessment in the national (Vögeli-Mantovani, 1999) and in the international literature (Brown et al., 2004). That is an important prerequisite for the implementation of such methods on a broader level (Clarke & Hollingsworth, 2002).

8.3.2 Benefits and challenges of assessment methods as mentioned by the teachers

Summary of overall results (copied from the last part of 7.3.2)

In the study, three formative assessment methods were trialled. From the teachers' evaluations of the benefits and the challenges of the different assessment methods, ten categories, covering the quotes for both school levels, emerged: Embedding formal formative assessment methods in inquiry-based science education; diagnosis of students' levels of achievement; content of feedback; role of the teacher; use of the feedback; learning effects; social and motivational effects; documentation; relation between formative and summative assessment; effort needed. Some of the categories were used to code either benefits or challenges, others were used for both.

The frequency of the different categories mentioned across methods was analysed. Considering the benefits, the content of the feedback as well as the learning effects were mentioned most frequently at primary school. At upper secondary school, the learning effects were mentioned most frequently. Considering the challenges, the diagnosis of the student levels and the content of the feedback as well as the effort needed being the most frequently mentioned by the primary school teachers and the same sub-categories plus the challenges associated with the embedding of formal formative assessment being the most frequently mentioned by the upper secondary school teachers.

Comparison of overall results to the literature

The occurrence of the same themes in the quotes on benefits and challenges can be taken as a sign that – across the two school levels – the aspects that are relevant for the teachers when trying out a new formative assessment method remain the same. The frequently mentioned categories were considered relevant by many teachers in the study.

Challenges related to designing the assessment activities, which are anticipated in the literature (Cizek, 2010; Swaffield, 2008; Yin et al., 2008) were not brought up by the teachers: Nobody mentioned difficulties in formulating assessment criteria, in finding an appropriate artefact to diagnose student learning, or in diagnosing student learning as a teacher. A possible interpretation is that teachers are used to these activities from summative assessment or that they appeared, compared to other activities, simply not very difficult.

Summary of results on specific assessment methods (copied from the last part of 7.3.2)

Speaking about written teacher assessment, the teachers from both school levels mentioned the advantages related to the quality of the diagnosis and the respective feedback to the students which was expected to lead to student learning in terms of scientific concepts and transversal competences, but also to an effect in the relation between teachers and students and an effect on student motivation. On the other hand, the definition of assessment criteria beforehand; the limited amount and extension of feedback; the doubtful use of the feedback by the students; the check-like character of written teacher assessment; and the big effort in terms of time were mentioned as challenges of written teacher assessment.

On peer-assessment, the teachers in the study mentioned the following advantages: The quality of the feedback in terms of language and its acceptance due to the fact that the assessor is a peer; the responsibility for the learning which lies with the students, resulting in a lower workload for the teacher and a higher capacity for individual support; learning effects in terms of transversal competences (communication etc.) as well as effects on the classroom climate and the students' motivation. Lastly, the low preparation time for the teacher was mentioned. Considering the challenges, the teachers from upper secondary school mentioned difficulties related to the planning of peer-assessment activities. Furthermore, teachers from both school levels expressed their doubts about the quality of the diagnosis and the feedback provided by peers and their uncertainty about their own role. Peer-assessment was also considered rather time-intensive and dependent on a good training of the students.

Speaking about self-assessment, the teachers stressed the advantages related to the role of the teacher who has time for individual support. Furthermore, effects on the students' transversal competences and on their self-regulated learning were anticipated. Considering the challenges, the teachers uttered their uncertainty on the quality of the students' reflections and the time such reflections take, similarly to the peer-assessment method.

Comparison of results on specific assessment methods to the literature

In the literature, the advantages related to written teacher assessment are located in the quality of the diagnosis (Jonsson, 2014; Luft, 1999; Moskal, 2003) and the feedback (Nunes, 2004; Santos & Dias, 2006; Stracke & Kumar, 2010), so similar to this study. The effects are seen in the improvement at the task level (Darling-Hammond et al., 1995; Ni, 1997; Wiggins, 1998) rather than in terms of student-teacher relation or student motivation as in this study, however. Considering the challenges, an earlier study by Bruno and Santos (2010) finds results similar to this study in terms of the pre-defined criteria and their static nature as well as the content and the extension of the feedback. The use of the feedback which depends on the approaches of the individual student has also been previously identified (Hyland, 1998). Similar to the teachers in this study, both Bailey and Garner (2010) and Tuck (2012) find it challenging for teachers to combine their two roles as providers of both formative and summative assessment. The time-consuming nature of written teacher assessment, finally, has been previously described by Bharuthram (2015) and Luft (1999).

Overall, the teachers in this study find benefits and challenges related to the use of written teacher assessment which are similar to what is reported in the research literature, apart from the expectations of what effects the assessment will provoke.

Comparing the benefits and challenges of peer-assessment as mentioned by the teachers in the study to the literature, a number of aspects are similar. The specific language characteristics of feedback formulated by peers and the responsibility for learning have been previously reported in Black et al. (2004). No references on the resulting capacities of the teachers were found in the research literature, however. The effects of peer-assessment on the students' transversal competences (Topping, 2010) and on self-regulated learning (Hanrahan & Isaacs, 2001; Lin et al., 2001; Topping, 1998; Topping, 2010) have also been previously mentioned but not the effects on the classroom climate and on the students' motivation as anticipated by the teachers in this study. The preparation time was not covered in the literature either. Considering the challenges, the planning issues as brought up by the teachers in this study are not mentioned in the literature.

The quality of the diagnosis (Topping et al., 2000; Topping, 2010) and the quality of the feedback (Black et al., 2003) have been previously discussed. The uncertainty about the own role that resulted, according to the teachers in this study, from the questionable quality of the diagnosis and the feedback, was not found in the literature. The lesson time and the training needed were recognized by Topping (2010), too. None of the teachers in the study spoke about the difficulties in what feedback to use for revision as reported in Sluijsmans (2002).

Overall, the benefits of peer-assessment perceived by the teachers in this study are similar to what is mentioned in the research literature. As in the case of the written teacher assessment, the social and motivational benefits from peer-assessment have not been found in the literature, though. The challenges of peer-assessment in the literature were not specifically focussed on the perspective of the teachers and their role nor on organisational issues, resulting in a smaller congruence between the results of this study and the research literature.

Consistent with the teachers' quotes on self-assessment in this study, the research literature suggests positive effects on the students' transversal competences (Boud, 1990; Harvey & Knight, 1996; Kwan & Leung, 1996; Pintrich, 2000; Zimmerman & Schunk, 2004) and on the self-regulation of the students (Boekaerts et al., 200; Nicol & Macfarlane – Dick, 2006; Schunk, 2003; Zimmermann & Schunk, 2002). The benefits for the teachers as perceived in this study were not found in the literature. Instead, a number of authors (Andrade et al., 2008; Evans, 2001; Hart, 1999; Wilcox, 1997; Yancey, 1998) report positive effects of self-assessment on the students' learning and achievement which were not mentioned by the teachers in this study. The doubtful quality of the students' self-assessment and the lesson time needed (Hanrahan & Isaacs, 2001) have been previously identified in the literature. Furthermore, the provision of clear assessment criteria and the practice needed by the students for meaningful reflections were also brought up by the same authors (Hanrahan & Isaacs, 2001) but not in this study.

Overall, the benefits and challenges of self-assessment mentioned in the literature focussed on the students' perspective rather than on the role of the teacher. Similar to the two assessment methods previously discussed, discrepancies between the research literature and this study were found in the effects of self-assessment.

Emerging differences between the results of this study and the literature

Comparing the benefits as perceived by the teachers across the assessment methods to the effects of formative assessment as proposed in the literature, a fundamental difference emerges: Social and motivational aspects which were prominently mentioned in all assessment methods in this study are hardly covered in the research literature. Instead, effects on student achievement are usually researched as the main benefit of formative assessment (see, for example, the meta-studies by Black and Wiliam (1998) or Natriello (1987), and sub-chapter 3.3). Interdependencies between formative assessment and student motivation (Black & Wiliam, 1998) and a relation between formative assessment and student confidence (Smit, 2009) have been suggested, but literature on these effects is generally scarce.

Comparing the results of the teachers from the two school levels

Overall, the teachers from the two school levels mentioned benefits and challenges that fell into the same categories. This can be taken as a sign that the aspects which are relevant to the teachers when speaking about formative assessment remain the same across school levels.

Looking at the sub-categories within the different categories in more detail, differences between the two school levels appear. Speaking about the benefits of written teacher assessment, the primary school teachers found that pre-defined assessment criteria were useful when collaborating with a team-teaching partner at the same class (

Table 27, sub-category 'pre-defined criteria'). The teachers from upper secondary school did not mention this aspect, which may be a hint for the occurrence of team-teaching at the different school levels. Talking about the value of written feedback, the teachers from primary school mentioned that it helped the students to focus on the learning goals and to plan their further learning (

Table 27, category 'content of feedback') stressed its role as a means of communication with the parents (

Table 27, category 'documentation'). At upper secondary school, the value of written feedback was found in its endurance compared to oral feedback which was considered more volatile.

Considering the challenges related to written teacher assessment, the primary school teachers were concerned about the students' use of the feedback provided (

Table 28, category 'use of feedback'): They mentioned that some students were resistant to feedback or that they may not understand it. These concerns were not uttered by the teachers from upper secondary school. Instead, these teachers felt unsure about the extent of feedback that should be provided since it might interfere with the openness of the inquiry (

Table 28, category 'content of feedback'). These differences could reflect the different approaches the teachers from the two different school levels have to student learning: Whereas the primary school teachers might aim at all students reaching a certain level of proficiency and therefore provide as much help as they think is necessary for every student to reach these minimal standards, the teachers from upper secondary school might expect their students to work more independently.

Speaking about peer-assessment, the perceptions of the benefits differed. Whereas the teachers from primary school mentioned that the feedback comes timely and is easy to understand because of the familiar language and vocabulary, the teachers from upper secondary school found that the value of the feedback lay in the fact that it was easy to accept and taken serious because the assessors were peers. The later also anticipated that feedback from peers might provoke further discussions as the inhibition level to do so was lower with peers than with the teacher (Table 29, category 'content of feedback'). The differences between the school levels could imply that self-regulated learning is considered important at upper secondary school but not so much at primary school. Further evidence that supports this interpretation can be found in the quotes on the role of the teacher where the teachers from upper secondary school mentioned that in peer-assessment, the students took the responsibility for their own learning (Table 29, sub-category 'role of the teacher') as well as in the quotes on social and motivational effects, where the upper secondary school teachers perceived peer-assessment as a way to take students serious and also an opportunity for them to show their respect towards the other students (Table 29, category 'social and motivational effects'). None of these aspects were mentioned by the primary school teachers. Differences occurred also in the teachers' expectations of what the students would learn through peer-assessment: Whereas the social development, communication to peers and an enhanced feedback culture as well as improved reflection was mentioned at both school levels, only the upper secondary school teachers expected their students to improve in their understanding of scientific concepts and in their inquiry competences (Table 29, category 'learning effects'). This could be seen as another fundamental difference between the teachers from the two school levels.

Looking at the challenges of peer-assessment, the teachers from upper secondary school mentioned several planning aspects that fell into the embedding of peer-assessment which were not brought up by the primary school teachers (Table 30, category 'embedding'). Some of these aspects can be related to the coordination difficulties that may occur when several teachers work with the same class (which is the fact at upper secondary school but less so at primary school). Other aspects could be interpreted as signs that the upper secondary school teachers are less used to the students working at different speeds than the teachers from primary school. Differences were also found in the perception of challenges related to the content of the feedback: Whereas the teachers at primary school worried that the students could be unable to follow feedback rules, the teachers from upper secondary school felt that some students were more critical than others, that the feedback from peers was not very reliable because of the content knowledge of the students (Table 30, category 'content of the feedback'). Consequently, the teachers from upper secondary school were unsure about when and how to interfere in case of mistakes (Table 30, category 'role of the teacher'). Finally, only the upper secondary school teachers mentioned that feedback from peers was often applied without reflection about its validity (Table 30, category 'use of feedback').

The two school levels were not compared for self-assessment because of the small size of the sub-samples.

Emerging differences between school levels in the approaches to formative assessment

Overall, two fundamental differences between the two school levels emerge in the approaches to the formative assessment methods: The underlying aim of formative assessment, for the primary school teachers, seems to be to support all their students in reaching a certain level of performance. For them, it is for example relevant that written teacher assessment helps the students to focus on the learning goals and to draw conclusions on their next steps in learning and they worry about students not making use of the feedback received. At upper secondary school, the underlying aim of written teacher assessment appears to be the autonomy of the students: The teachers worried, for example, about the interferences between feedback and the student-oriented nature of inquiry. This first dichotomy becomes more predominant when the teachers speak about the advantages and challenges of peer-assessment (see above paragraph on the advantages of peer-assessment).

Looking at the specific values and mechanisms of the different assessment methods, no fundamental differences between the teachers of the two school levels were found for written teacher assessment and for self-assessment. But the main value of peer-assessment, for the primary school teachers, was the students' learning in social and communicational competences as well as in their reflective abilities. At upper secondary school, these learning effects were also mentioned. But additionally, the teachers from upper secondary school also mentioned benefits in the understanding of scientific concepts and inquiry-competences (which the peer-assessment was supposed to be targeted to). They therefore worried about the validity of peer-assessment which the teachers from primary school did not. It can be interpreted that the primary school teachers expect their students to benefit from providing feedback whereas the upper secondary school teachers also hope their students to make progress based on the feedback they receive from their peers.

8.3.3 Means of support as mentioned by the teachers

Summary of results (copied from the last part of 7.3.3)

The teachers from primary and from upper secondary school level participating in the study were asked about possible means of support for formative assessment in their classrooms. The teachers' answers were similar across school levels and will therefore be summarized without reference to those. In total, seven means of supports were mentioned: Provision of examples of good practice; time (e.g. for planning the activities; for providing feedback; etc.); support from team-teaching partner or another person with a teacher-like function; training and coaching to enhance the teachers' assessment literacy; opportunities and prompts to reflect upon assessment practices; platform to exchange experiences and problems with peer teachers; and clarification of the role of formative assessment and its relation to summative assessment at the level of educational policy.

Comparison to challenges

Comparing the measures of support suggested by the teachers to the challenges from section 8.3.2, it appears that most of the challenges could be approached with different measures of support suggested: The provision of examples of good practice; the enhanced assessment literacy, the reflection upon assessment practices and the exchange of experiences and co-construction of knowledge on assessment amongst peer-teachers could all feed into overcoming several challenges mentioned in the first part of the results, namely the embedding of formative assessment into a unit; the content and the structure of the feedback; the students' use of the feedback, and, to some extent, also the effort needed. Time as a means of support matches the effort needed on the challenge side. The clarifications of the assessment policy will tackle the unclear relation between formative and summative assessment.

Overall, the teachers' perceptions of challenges with the different assessment methods in classroom appear consistent with the measures of support on different levels they suggest.

8.3.4 Usability as perceived by the students

Summary of results (copied from the last part of 7.3.4)

Selected classes with upper secondary school were asked about the usability of peer-assessment from their perspective. Generally, the students perceived both the role as assessors and assesses positively. Reasons for these evaluations can be found in the subsequent sections.

Implications and comparison to the literature

This can be taken as a sign that peer-assessment as a formative assessment method is accepted by upper secondary school students. Furthermore, the students seem to be able to recognize advantages not only related to acting as assessors but also related to the feedback received. This has also been suggested by Hanrahan & Isaacs (2001); Lin et al. (2001); Topping (1998; 2010), and it is consistent with the results from the teachers (see 8.3.2; section 'emerging differences in the approaches to formative assessment'). The students' perceptions of formative assessment are relevant because the students' effort and engagement in the classroom is likely to influence the success of lessons containing formative assessment activities.

8.3.5 Benefits and challenges as mentioned by the students

Summary of results (copied from the last part of 7.3.5)

In the study, students from five upper secondary school classes were asked about the benefits and challenges of peer-assessment they perceived. The themes that emerged from the students' evaluations could be completely covered with a sub-set of the categories derived from the respective teacher answers. Speaking about the benefits, the students mentioned 'content of feedback', 'learning effects', and 'social and motivational effects'. Speaking about the challenges, the students mentioned 'diagnosis of students' levels of achievement', 'content of feedback', 'social and motivational effects', and 'effort needed'.

Implications and comparison to the teacher results

The categories that summarize the benefits and challenges of peer-assessment according to the students at upper secondary school are a sub-set of the categories mentioned by the teachers. The underlying aims of peer-assessment that were identified from the teacher answers (see 8.3.2), that students should develop in their self-regulation and that they should be able to improve their work from peer-assessment (not solely their communication and social abilities) were found in the student answers as well. The result can be taken as a sign that students and teachers at upper secondary school agree to a great extent on their evaluations of the assessment method.

The students focussed on the quality of the diagnosis and the feedback as well as on the effect on learning and on motivation. Planning issues (embedding the methods), the role of the teacher, the use of the feedback, the documentation and the relation between formative and summative assessment were not brought up. Some of these aspects may not be relevant to students (planning issues and the role of the teacher) whereas in other cases, the students did not seem to be aware of challenges: The use of feedback and its documentation was considered a problem by the teachers but not by the students. In order to reduce the teachers' frustration with formative assessment in their classroom, the students should be sensitized for the documentation and use of the feedback.

Considering the sub-categories, the students' answers are a sub-set of the system developed from the teachers' answers. The exception is the expected effects of peer-assessment on student learning: The teachers did not anticipate the students to learn on the nature of science whereas the students did.

8.3.6 Means of support as mentioned by the students

Summary of results (copied from the last part of 7.3.6)

In the study, students from five upper secondary school classes were asked about possible means of support for formative peer-assessment in their classrooms. The students' answers fell into six categories: (1) no support needed; (2) support in formulating feedback; (3) structuring questions or criteria to focus on; (4)

anonymity; (5) access to content knowledge or to the correct solution; and (6) exchange with peers or with the teacher.

Comparison to the literature

Similar to the last section, the upper secondary students' answers direct to the concrete challenges of diagnosing and providing feedback itself rather than to planning and strategic issues. The results also show that the support falls into two large groups, one of them being on the student assessment literacy and the other one on their content knowledge. This implies that peer-assessment, firstly, has to be scaffolded and practiced (as suggested by Black et al., 2003; Topping et al., 2000). Secondly, it also signifies that students need, at least at upper secondary school, certainty about the 'correct' solution in the end. This issue has not been discussed extensively in the literature.

Comparison to the challenges mentioned

Comparing the means of support suggested by the upper secondary school students to the challenges with peer-assessment they perceived, the specific roles of the peers could be addressed by the structuring questions or criteria as well as the access to the correct solution and the exchange with peers. The formal and language-related issues could be tackled by both the supporting formulating feedback and the structuring questions. Anonymity could help to approach the motivational and social issues. The effort needed is difficult to bypass, no concrete solution was suggested by the students.

Overall, the means of support and the challenges perceived by the students have a high conformity.

Comparison to the teachers

Similar to the challenges of peer-assessment, the upper secondary school students' suggestions for measures of support can be considered a sub-set of the teachers' answers: The students naturally focussed on the quality of peer-assessment at classroom level whereas the teachers also targeted the level of educational policy and others.

8.4 Discussion of research question 4: Changes in teachers' understandings and implementations throughout the collaboration in the study

In this sub-chapter, the results of research question 4 '*How do the teachers' understandings and implementations of formative assessment change throughout the study?*' will be summarized, discussed and compared to the literature.

With the small sample sizes, the results on research question 4 are clearly tenuous. Due to the little literature available on changes in teachers' formative assessment practices and beliefs throughout the collaboration in a project where the teachers develop their own assessment, it nevertheless appeared legitimate to conduct the respective analyses. The interpretation of the results will be done conservatively. Part of this cautious interpretation is that the data was, in some sections, not analysed separately for the two school levels as for the other research questions.

8.4.1 Changes in the teachers' descriptions of what formative assessment is throughout the collaboration in the study

Summary of results (copied from the last part of 7.4.1)

The teachers' descriptions of what formative assessment and their changes throughout the collaboration (beginning of collaboration; middle phase; end of collaboration) in the study were analysed. The 11 categories developed to code the teachers' answers for research question 1 were used again: Five categories describe elements that are also present in the literature on formative assessment; one element is generally ascribed to assessment in the literature; four categories describe other elements and one category contains examples of assessment methods. The general analysis shows that the teachers' descriptions of formative assessment generally converged towards what can also be found in the literature (supportive nature; guidance for next steps in learning or teaching; individual; prospective rather than retrospective). The analysis throughout the time of collaboration shows that four groups of teachers can be made: The first group in which the teachers' understanding of formative assessment did not change throughout the study (seven teachers); the second group in which the teachers improved towards descriptions which can also be found in the literature (five teachers); the third group in which the teachers' descriptions worsened (the opposite direction than the second group; one teacher); and unclear directions (two teachers).

Implications and comparison to the literature

The results imply that teachers' description of formative assessment can change throughout the collaboration in a study and that a misconception can disappear. For the teachers of the study, their descriptions did not just change in any direction but either improved or they held the level in almost all cases. This is in congruence with an earlier study on the effects of a collaborative study with teachers by Marshall and Drummond (2006) saying that the teachers' beliefs on assessment are not stable but may be influenced by the collaboration in a project. The finding is also in accordance with the model of professional growth (Clarke & Hollingsworth, 2002, see sub-chapter 4.1) which suggests that teacher knowledge, beliefs and attitudes interact with the professional experimentation and with the experience on the outcomes of such experimentation, but also with external sources of information or stimuli.

8.4.2 Changes in the self-efficacy

Summary of the results (copied from the last part of 7.4.2)

The change in the teachers' personal formative assessment efficacy belief was measured by the respective items in the teacher profile questionnaire at the beginning and at the end of collaboration in the study. The results show that the collaborating teachers' personal formative assessment efficacy belief increased significantly whereas the control groups' efficacy did not change. The control group consisted of peer teachers who did not collaborate in the study.

Comparison to the literature and implications

In the literature on teacher self-efficacy (see section 3.5.2), it is confirmed that self-efficacy can change (Ash-ton & Webb, 1986; Ramey-Gasset & Shroyer, 1986). Four contributors leading to this effect have been identified (Bandura, 1977; 1982): Mastery experience (which could be linked to the practical experience with the formative assessment methods in this study); vicarious experiences and social persuasion (which could both be linked to the interaction and reflection with peer teachers in this study); as well as physical and emotional factors which are strongly dependent on the particular situation and person involved and therefore cannot be linked to the study. Overall, different factors may have contributed to the enhanced personal formative assessment efficacy belief in this study.

8.4.3 Changes in the trials throughout the collaboration in the study

Summary of the results (copied from the last part of 7.4.3)

The quality and the quantity of the formative assessment activities trialled throughout the study (in three rounds) were analysed. The resulting picture is not very clear at the level of the individual teacher. However, what can be said is that in the first round, almost all trials conducted followed the criteria as defined in sub-chapter 5.4 and were successful in that sense. In the second round, a considerable number (five trials out of 16) did not fulfil all criteria defined in 5.4. In the third round, most of the primary school teachers either left the study or did not conduct any trials.

In order to analyse the quantity of formative assessment activities, the respective items from the teacher profile questionnaire were analysed. No significant differences between the beginning and the end of the study were measured.

Implications and comparison to the literature

A possible explanation for the unclear emerging picture in terms of implementation quantity could be that it was never the aim to conduct as many formative assessment activities as possible. Instead, the teachers were supposed to trial one formal method per semester.

Looking at the implementation quality, a possible explanation for the results is that in the first round of implementation, the teachers from both school levels conducted trials as requested. Since they had the choice to select the method and specific procedures themselves (see sub-chapter 5.1) and as the implementation of formative assessment methods is reported to be challenging (e.g. Furtak et al., 2008), it could be that many of the teachers tried out safe procedures. 'Safe' in this context means that the teachers chose procedures which they felt sure would function. In the later rounds, the teachers from primary school either became more imaginative, trying out more risky strategies that did not follow the criteria given in the study, or they did not take the time for trials anymore. Similarly, the British study introduced in sub-chapter 3.7 (Black et al., 2003; Wiliam et al., 2004) found a high variability between teachers in the quality of the activities implemented. At upper secondary school level, the teachers continued to conduct trials as requested by the guidelines of the study.

The differences in the behaviour of the teachers from the different school levels could be related to the differences in their socialisation; with primary school teachers generally being more creative and innovative and upper secondary school teachers being more straight-forward but also more reliable. In that sense, it is possible that the open setting of the study provoked some teachers to use the room for experimentation rather than to conduct safe trials in the later stages of collaboration.

8.4.4 Changes in the importance, benefits and challenges of the formative assessment methods perceived by the teachers throughout the collaboration in the study

Summary of the results (copied from the last part of 7.4.4)

Changes in the teachers' perception of importance of formative assessment and challenges and benefits of the formative assessment methods throughout the study were explored. The results show that the perceived importance of formative assessment did not change. No pattern or tendency could be found in the advantages and challenges mentioned either.

Implications

Both results could imply that the teachers' perception of benefits and challenges did indeed not change; that the sample sizes were too small to see any such pattern; or that the coding system applied was not suitable to investigate the question. No literature covering the topic could be found.

8.4.5 Support mechanisms from the collaboration in the study

Summary of results (copied from the last part of 7.4.5)

The teachers collaborating in the study were asked what aspect(s) of the study they considered helpful: The theoretical background information on formative assessment; the provision of concrete ideas and examples to draw from; the opportunity and the prompts to try out different methods; the opportunity and prompts to reflect upon the formative assessment practices; the exchange with peer teachers; and the broadening of the horizon through the contact with teachers from a different school level.

Comparison to the literature

The provision of theoretical background information and concrete ideas and examples can be related to the literature on classical professional development programmes (e.g. Brookhart et al., 2010; Mertler, 2009; Sato et al., 2008). However, the transfer of the knowledge from the professional development program to teaching practice is reported to be difficult (e.g. Maier, 2015).

The exchange with other teachers which was prominently mentioned in this study can be related to the effects of professional learning communities (PLC) as reported in Fulton & Britton, 2010 (see section 3.5.2). The two authors describe that the work in professional learning communities, amongst other effects, engaged teachers in discussion about pedagogical strategies.

The prompts to try out formative assessment methods and to reflect upon the formative assessment practices can potentially be linked to the literature on more innovative forms of teacher professional development. These include the teachers developing their own assessment (see section 3.7.2) and school development projects (see section 3.7.3). However, neither of the authors (Black et al., 2003; Wiliam et al., 2004; Smit, 2008) investigating these innovative forms of professional development explored the mechanisms that lead to the success in their programmes.

The broadening of the horizon through the contact with teachers from other school levels that was mentioned by teachers in this study has not been covered in the literature on the implementation of formative assessment. Even though this answer was potentially triggered by the specific setting of this study, this field of research may deserve more attention.

8.4.6 Variability of implementations within teachers

Summary of results (copied from the last part of 7.4.6)

In their implementation habits, differences between different teachers in the study were found: Whereas some teachers had high variabilities between their trials, others had a high overlap between one trial and the subsequent one. The extent of the variability remained stable for the individual teachers across the three rounds of implementation. Despite intermediate and large effect sizes, no significant differences were found between the teachers of the different school levels. At the level of gender, significant differences were found

in the variability of assessment methods: The female teachers collaborating in the study varied the assessment methods trialled significantly more than the male teachers. The respective effect size was large. Significant differences were also found between the teachers with less than ten years of teaching experience and the teachers with a higher teaching experience in the study: The more experienced teachers varied their inquiry contexts significantly more than their less experienced peers. The effect sizes were large again. The last results, however, depend on the formation of the subgroups: If, for example, the less experienced teachers were defined as 0-4 years of teaching experience and the more experienced teachers as >4 years of teaching experience, the differences disappeared.

Implications

The results show that the implementation behaviour of the different teachers in the study, measured by the variability of their implementations, depends on different variables. School level, teacher gender, and teaching experience might all have an influence on different aspects of the implementations. The medium- and large Cohen effect sizes can be taken as a sign that the influences of the different variables are relevant even though it is difficult to find significant results with the small sample sizes.

Comparison to the literature

A British study described in Black et al. (2003) and Wiliam et al. (2004) also investigated the implementation behaviours amongst their teachers. They distinguish between four different implementer types (Black et al., 2003, see 3.7.2) based on the use of formative assessment strategies which roughly correspond to different formative assessment methods: Trailers, static pioneers, moving pioneers, and experts.

These results are consistent with the study here in the sense that both studies find different implementation behaviours amongst teachers. In the context of this study, it was, however, not possible to find four implementer types: The study reported in Black et al. (2003) and Wiliam et al. (2004) did not only investigate the variability of subsequent implementations but also measured the success of the teachers' implementations by calculating their effects on student achievement. This success fed, along with the sustainability of the use, into the definitions of the different implementer types. Success and sustainability of the use of the methods was not measured in this study here.

These differences in the design could explain why the results from this study on the variability of subsequent implementations do not suggest different types of teachers: The overlaps do not group but seem to be distributed on a continuum between two extremes. One extreme describes conservative teachers who did not change their approaches from one trial to the next trial. They could be related to the 'static pioneers' from the British study ("teachers who were successful with one or two key strategies and who had restricted themselves to these"; Black et al., 2003, p. 28). The other extreme describes very innovative teachers who trialled completely different approaches. They could be related to either the 'moving pioneers' ("teachers who were successful with one or two key strategies, but having routinized these were looking for other ways to augment their practice"; Black et al., 2003, p. 28), or to the 'experts' who have "formative assessment strategies embedded in and integrated in practice" (Black et al., 2003, p. 28). But the data of this study does not allow for a stronger connection between the two studies.

8.5 Measures of support for formal formative assessment practices in Switzerland

In this sub-chapter, the first of two aspects that overarch the four research questions of this study will be discussed: The measures of support that might enhance formative assessment practices in inquiry-based science education in Switzerland.

The sub-chapter is structured into six sections that cover teacher concepts of formative assessment which appear present at the two school levels (8.5.1); the general attitude towards formative assessment (8.5.2 and 8.5.3); formative assessment practices that appear realistic at the school levels explored (section 8.5.4); problem areas where support is needed (8.5.5) and finally support for the uptake of formative assessment (section 8.5.6). Every section will conclude with a hypothesis on the implementation of formal formative assessment methods in inquiry-based science education in Switzerland.

8.5.1 Teacher concepts and misconceptions of formative assessment

When trying to support the implementation of formative assessment in teaching practice, it is important to consider what teachers already know about formative assessment. The results on research question 1 (see sub-chapters 7.1 and 8.1) can be taken as a first insight to such knowledge. A portion of teachers in the study from both primary and upper secondary school offered an explanation of the term ‘formative assessment’ that was similar to what can be found in the literature. But three misconceptions could also be identified: Firstly, that formative assessment was focussed on a specific type of competences or goals such as social abilities or other so-called soft skills; secondly, that formative assessment has an individual reference norm; and thirdly, that formative assessment is grading of the learning process (instead of grading of the product which would correspond to summative assessment).

These results are taken as a basis for hypothesis H1: Apart from the concept which can also be found in the literature (formative assessment as a means to support student learning and teaching and therefore having a prospective orientation), a number of misconceptions on formative assessment exist amongst teachers from different school levels in Switzerland.

8.5.2 Teacher attitude towards formative assessment

The conditions for an uptake of more formal formative assessment in daily teaching practice appear, from the data of the study, to be generally positive for three reasons: Firstly, the teachers from the two school levels collaborating in the study were able to integrate formative assessment methods meaningfully in their inquiry teaching (see results on research question 2). Secondly, the teachers from both school levels in the study considered the formative assessment methods trialled usable at their school level to a great majority – and so did the students at upper secondary school (see results on research questions 3.1 and 3.4). Thirdly, the teachers from both school levels who collaborated in the study recognized a number of concrete benefits from the formative assessment methods trialled for their classrooms – and so did the students at upper secondary school (see results on research questions 3.2. and 3.5). The students are considered relevant because their acceptance of the formative assessment methods might heavily influence the success of respective lessons. And in the Swiss educational system, where the teachers’ autonomy in designing their teaching at both school levels is high, the teachers’ and the students’ acceptance of an innovative feature is central for its implementations.

These three pieces of evidence are taken as a basis for hypothesis H2: Not only the teachers collaborating in the study but teachers in Switzerland generally have a positive attitude towards using formal formative assessment methods in their inquiry-based science education.

8.5.3 Aims pursued with formative assessment

Apart from getting an insight into teacher knowledge on formative assessment, the results from this study also allow for conclusions on the aims the teachers from the two school levels pursue with formative assessment activities. Such considerations may help to convince a broader community of teachers why they should use formative assessment strategies in their teaching.

From the benefits of the three formative assessment methods trialled, commonalities and differences in the underlying aims of formative assessment of the teachers at the two school levels (see results in 7.3.2 and discussion in 8.3.2) have been found and will be summarized in Table 55.

Table 55: Approaches to formal formative assessment at the two school levels investigated.

	Approaches to formal formative assessment of the primary school teachers in the study		Approaches to formal formative assessment of upper secondary school teachers in the study
Aims of formal formative assessment	Formative assessment as a means to ensure that all students reach a certain level of performance (see 8.3.2). Formative assessment as a means to enhance student motivation, the student-teacher-relation, and the classroom climate (see 8.3.2)	↔ =	Formative assessment as a means to enhance the students autonomy and their self-regulated learning (see 8.3.2) Formative assessment as a means to enhance student motivation, the student-teacher-relation, and the classroom climate (see 8.3.2)
Mechanisms through which students learn and learning effect of written teacher assessment	Written teacher assessment as a means to foster conceptual understanding and transversal competences. The benefit lies in the clarification of expectations, in the careful diagnosis, in the nuanced feedback (see Table 27 and discussion in 8.3.2)	=	Written teacher assessment as a means to foster conceptual understanding and transversal competences. The benefit lies in the clarification of expectations, in the careful diagnosis, in the nuanced feedback (see Table 27 and discussion in 8.3.2)
Mechanisms through which students learn and learning effect of peer-assessment	Peer-assessment as a means to foster the students social, communicational, and reflective competences: The benefit lies in the provision of feedback (see Table 29 and discussion in 8.3.2)	↔	Peer-assessment as a means to foster the students social, communicational, and reflective competences but also their conceptual understanding and science-specific competences: The benefit lies in both the provision but also the use of the feedback (see Table 29 and discussion in 8.3.2)
Mechanisms through which students learn and learning effect of self-assessment	No hypothesis possible based on the limited amount of data		

These approaches to formative assessment and respective methods (see Table 55) from enthusiastic and innovative teachers are taken as a basis to model teacher aims with formative assessment for the two school levels explored. Knowledge on these aims might help to convince teachers to embed formal formative assessment in their teaching. Whereas teachers from primary school aim at helping all students to reach a certain level of performance, teachers from upper secondary school try to enhance the students' autonomy

and self-regulated learning. Further differences exist between the concrete methods of formative assessment. These contrasts reflect the disparity between the more pedagogic approaches of Swiss primary school teachers who aim at holistically educate their children and the approaches more focussed to teach subject-specific knowledge of upper secondary school teachers. But apart from the effects of formative assessment on student learning, the teachers from both school levels anticipate social and motivational effects from formative assessment which are barely covered in the research literature.

The data from the study and the subsequent considerations are taken as a basis for hypothesis H3: The aims in terms of student learning which teachers pursue with formal formative assessment methods differ between school levels. Apart from the student learning, the teachers from the two school levels explored also aim at provoking motivational and social effects through formative assessment.

8.5.4 Formative assessment practices for the school levels explored

The data on the trials in the study suggest that there are no clear rules on what combinations of assessment methods and competences lead to most success. Many approaches have been put into practice meaningfully, and a long list of variables (including teacher personality; classroom climate and character of the student group; time available; and many others) seem to influence the trials. The variability in the trials reflects the teachers' autonomy in designing their teaching at both school levels and the far-reaching responsibilities which are connected to this autonomy.

However, differences between the trials at the two school levels (see results in 7.2 and discussion in 8.2) have been found and will be summarized in Table 56. These contrasts reflect the disparity between the more pedagogic approaches of Swiss primary school teachers who aim at holistically educate their children and the approaches more focussed to teach subject-specific knowledge of upper secondary school teachers.

These characteristic features (see Table 56) from the trials of enthusiastic and innovative teachers are taken as a basis to model types of formative assessment practices that appear feasible for the two school levels explored. At primary school level in Switzerland, it is anticipated that more complex formative assessment practices are possible such as the combination of several methods within the context of one inquiry unit. At upper secondary school level in Switzerland, it is anticipated that realistic formative assessment will be rather efficient, simple and closely linked to summative assessment practices.

Table 56: Characteristic features of the trials in the study

	Model of formative assessment practices at primary school	↔	Model of formative assessment practices at upper secondary school (<i>Maturaarbeiten</i> not considered)
Inquiry characteristics (see 7.2.1)	Rather open inquiries (see Figure 12)	↔	Inquiries pre-defined to a higher degree (see Figure 12)
	Smaller number of inquiry activities (see Figure 14)	↔	Higher number of inquiry activities (see Figure 14)
	Focus on planning and conducting (see Figure 13)	↔	Focus on planning, conducting, analysing/interpreting and communicating (see Figure 13)
Formative assessment characteristics (see 7.2.2)	Several assessment method-competence combinations in one unit (see Figure 22; Figure 23)	↔	One assessment method-competence combination in one unit (see Figure 22; Figure 23)
	Formative assessment based on various sources (written artefacts; observations, ...) with observational data being the most common (see Figure 20; Figure 21)	↔	Formative assessment based on one source of data: written student data (see Figure 20; Figure 21)
	Formative assessment focussed on conducting inquiries (see Figure 15)	↔	Formative assessment focussed on documenting inquiries (see Figure 15)
	Transversal competences play an important role (see Figure 18)	↔	Transversal competences play a minor role (see Figure 18)
	Feedback is used to revise the original artefact in short cycles (see Figure 24; Figure 26)	↔	Feedback is transferred to subsequent tasks in longer cycles (see Figure 24; Figure 26)

The data from the study and the subsequent considerations are taken as a basis for hypothesis H4: Formative assessment practices in the context of inquiry-based science education that could realistically be expected from a considerable portion of teachers differ between different school levels.

8.5.5 Problem areas for the uptake of formative assessment practices from a teacher perspective

A number of the sections in the result chapter provide evidence on what problems the teachers perceive in the uptake of formative assessment practices. These problems were grouped into four problem areas which are, naturally, intertwined to some extent: Teacher assessment literacy; use of lesson time; availability of resources; and the position of formative assessment within the assessment framework. Various data from the study will be laid out and taken as a basis to formulate a hypothesis on the areas where support for the uptake of formal formative assessment practices is needed.

Insights into the teacher assessment literacy are provided from the data on problems in the trials (see 7.2.3) the teachers in the study were or were not aware of. Secondly, there is data on the challenges associated to every assessment method the teachers perceived (see 7.3.2). The data on the problems in the trials (see 7.2.3) suggests that sharing assessment criteria with the students and providing an opportunity to use the feedback received, either for revision of the assessed piece of work or by transferring the feedback to a similar, subsequent activity, was difficult for some of the teachers from both school levels. The data from the teachers' evaluations of the formative assessment methods (see section 7.3.2) show that the teachers from both school levels need support in facing a number of method-specific challenges. These challenges

are situated on a practical level and include the embedding of the specific methods in a particular unit; the training and coaching the students need to conduct self- and peer-assessment; the classroom climate needed to provide usable peer-assessment; and efficient ways of providing formative assessment with a reasonable effort.

The problems related to the use of lesson time are reflected in the data on the challenges associated with the assessment methods as perceived by the teachers (see 7.3.2). The teachers clearly stated that they cannot dedicate unlimited lesson time to the introduction and student training in self-and peer-assessment, and that they cannot spend time to the provision of formative assessment unrestrictedly.

The availability of resources was mentioned as a challenge in the individual interviews (see 7.3.3). The teachers from both school levels said that they needed examples of formative assessment activities, prompting questions, and assessment criteria for direct application but also for inspiration.

Finally, the position of formative assessment within the assessment framework ways was also addressed in the individual interviews (see 7.3.3). At both school levels, the teachers in the study felt unsure about the exact meaning of the term 'formative assessment' and its delineation from summative assessment. They were also uncertain about the importance of formative assessment relative to summative tests and regional checks.

The data from the study and the subsequent considerations are taken as a basis for hypothesis H5: In Switzerland, the teachers from both school levels explore several problem areas regarding the implementation of formative assessment: Besides the teacher assessment literacy, the use of lesson time; the availability of resources; and the position of formative assessment within the assessment framework are also considered problematic.

8.5.6 Measures of support for the uptake of formative assessment practices

In result section 7.3.3, the teachers in the study suggested measures of support that could facilitate the uptake of formative assessment methods in their inquiry-based education. They will be discussed in the order of the problem areas from section 8.5.5.

The problems with teacher assessment literacy were, on the one hand, suggested to be approachable with classical pre- and in-service professional development (see 7.3.3). But the teachers from both school levels also stressed the importance of co-constructive approaches such as opportunities to exchange experiences on formative assessment with peer teachers and opportunities to reflect upon one self's assessment practices (see 7.3.3). These statements reflect the teachers' autonomy in their teaching at both school levels which is exceptional in Switzerland.

The use of lesson time for the introduction and student training in self-and peer-assessment could be approached by collaboration within schools: If teachers agreed upon what formative assessment strategies to introduce and to train at what grade and in what subject, time could be saved. A relief for the teacher, as suggested by a number of primary school teachers, could be achieved through team-teaching. This would help the teachers to gain time for individual support during classroom hours (see 7.3.3). The idea on collaboration within schools might be realisable with the school development projects becoming more and more common at all school levels. The second idea could be difficult because of the financial effort needed.

The availability of resources could, according to the teachers in the study, be improved by including examples of formative assessment activities in teaching resources such as school books (see 7.3.3). This suggestion is promising particularly at primary school where the new curriculum *Lehrplan 21* is being implemented: In the course of this implementation, new teaching resources are being developed with a focus on the development of student competences.

The position of formative assessment within the assessment framework should be clarified from the level of educational policy (see 7.3.3); with cantonal and national guidelines that take into account the school-level specific circumstances such as the curriculum, regional checks and summative exams. With the introduction of the competence-oriented curriculum *Lehrplan 21*, informative brochures are being produced and teacher in-service trainings, but also public discussions and similar are held. In the course of these activities, a stronger focus on assessment appears doable.

Considerations based on the data from the study are taken as a basis for hypothesis H6: Measures at different levels might support the uptake of formative assessment activities in the context of Switzerland with its high teacher autonomy and the new curriculum at the compulsory school levels: Classical pre- and in-service professional development; school development projects for the co-construction of knowledge on assessment and for the coordinated introduction of assessment strategies in classes; school books as a source of ideas for formative assessment activities; and a clear communication of the purpose of formative assessment in cantonal and national guidelines.

8.6 Teachers developing their own formative assessment practices in the context of this study

In this sub-chapter, the second out of two aspects that overarch the four research questions of this study will be discussed: Teachers developing their own formative assessment practices in the context of this study. The sub-chapter is structured into two sections: Effects of the study on the teachers' understandings and practices of formative assessment (section 8.6.1), and implementer types (section 8.6.2). Both sections will conclude with a hypothesis on the implementation of formal formative assessment methods in inquiry-based science education in Switzerland.

8.6.1 Effects of the study on the teachers' understandings and practices of formative assessment

The model of professional growth from Clarke & Hollingsworth, 2002, was described (see sub-chapter 4.1) and taken as a theoretical framework for the implementation of an innovative approach in the practice of teachers. The model consisted of four domains. The first one is the external domain with sources of information or stimulus. In this study, the external domain has both a regional or national horizon with the curricula (see theory part, sub-chapter 3.8), but also an international horizon with its OECD papers (see sub-chapter 3.6). The second domain, the personal domain from the model of Clarke & Hollingsworth, contains big constructs like knowledge, beliefs and attitudes. Aspects of the personal domain, the teachers' understanding of formative assessment (see sub-chapter 7.1 and 7.4.1) and their formative assessment self-efficacy (see section 7.4.2) were explored in the study. The third domain, the domain of practice, was covered with research question 2 (see sub-chapter 7.2 and section 7.4.3) and provides insights into the teachers' professional experimentation. The domain of consequences from the model from Clarke & Hollingsworth, finally, certainly include the effects of formative assessment on student learning as described in sub-chapter 3.3 – yet measuring the effect of different formative assessment practices on the students' achievement was never an aim of the study. The domain was, instead, investigated by asking students and teachers about the benefits and challenges of the different formative assessment methods in research question 3 (see sub-chapters 7.3 and section 7.4.4).

The model from Clarke & Hollingsworth does not only consist of domains but also of relations between these domains. It is distinguished between enactment and reflection. The relations are explored in research question 4 in this study. It is, however, hard to learn more about these relations. This becomes clear with the two following examples: Firstly, the results from sub-chapter 7.4.1 and 7.4.2 show that there are changes in the personal domain throughout the study but it is not possible to attribute these changes to a specific relation with any of the other domains. Secondly, there is anecdotal evidence in the data that (see section 7.4.3) the domain of practice and the domain of consequences are strongly intertwined but the exact nature of these interactions is hard to capture from the data available. So it is difficult to know what trigger exactly influenced the teachers' changes in their understanding and practices of formative assessment. Based on the data, it is only possible to ascribe the results from sub-chapter 7.4 to the collaboration in the study as a whole.

This collaboration was influenced by the considerations laid out in sub-chapter 3.7 and particularly by the articles on a collaborative project on formative assessment from Black et al. (2003) and Wiliam et al. (2004). That collaborative project basically followed the idea that there is not one ideal solution of how to formatively assess students for all teachers but that every teacher has to find his or her assessment practice with the help of programmes of professional development and support (see 3.7.2 for details).

In this study here, the work of the teachers was planned to be steered by four mechanisms (see section 5.1 for details): (1) The provision of a theoretical background and inspirational examples on formative assessment; (2) The urge to develop and trial formative assessment activities by asking every teacher to report upon one trial per semester; (3) Encouragement for collaboration between teachers with group discussions where formative assessment activities were exchanged and a dropbox that was used to exchange materials; (4) Encouragement for reflection triggered by the questions in the evaluation form and in the group discussions. As laid out in section 8.4.5, the teachers mentioned all of these mechanisms as helpful aspects of the

collaboration in the study. In addition, the broadening of the personal horizon through the contact with teachers from a different school level was also mentioned by some of the teachers.

Based on the results, it can be expected that the collaboration in the project as explained in the last paragraphs had an effect on the teachers' understanding of formative assessment (see sections 7.1 and 7.4.1) and on their formative assessment self-efficacy (see section 7.4.2). Changes in the formative assessment practices (quality or quantity; see section 7.4.3) could not be shown in the data.

The results from the study as summarized above are taken as a basis for hypothesis H7: The collaboration of teachers in a study on formative assessment with an open setting as described above, interpreted as an in-service training, can have an effect on the teacher's understanding of and attitude towards formative assessment. The effect could be provoked by different mechanisms: (1) The provision of a theoretical background and inspirational examples on formative assessment; (2) The urge to develop and trial formative assessment activities by asking every teacher to report upon one trial per semester; (3) Encouragement for collaboration between teachers with group discussions where formative assessment activities were exchanged and a dropbox that was used to exchange materials; (4) Encouragement for reflection triggered by the questions in the evaluation form and in the group discussions.

8.6.2 Implementer types

A large portion of the data from the study relates to the domain of practice from the model of professional growth. The data on the implementation stories of the individual teachers show (see section 7.4.6, variability of implementations within teachers) that there were different kinds of professional experimentation in the study.

According to the overlaps of the trials in terms of different variables (such as dimensions of openness, competences assessed, assessment methods, and others, see section 7.4.6), different implementation behaviours were found amongst the teachers in the study. Some teachers trialled different approaches in the three semesters whereas the other teachers developed their next trial based on the experiences of the last one, resulting in trials that were similar to each other (e.g. same assessment method; same inquiry context; same criteria or similar). The results of this study could be related to an earlier British study (Black et al., 2003; Wiliam et al., 2004) even though the results could not be exactly confirmed. The latter is potentially caused by the differences in the designs of the two studies.

The differences in the implementations behaviours between teachers could be influenced by different variables such as school level, gender, and teaching experience (see section 7.4.6 for details).

The results from the study as summarized above are taken as a basis for hypothesis H8: In the context of formative assessment practices, there are different implementation behaviours. Different variables (school level, gender, teaching experience) appear to have an influence on the implementation behaviour.

9 Retrospects and prospects

9.1 Aims of the study

Inquiry-based science education has been an important part of science educational theory and practice for the last decades. As in other approaches to teaching and learning, the appropriate support and assessment of the students' competences has been much debated in the context of inquiry-based science education.

One way to support and assess students in their learning is formative assessment. In this explorative study, trials of three formative assessment methods in inquiry-based science education at primary and at upper secondary school and the teacher perspective on those was investigated. Based on the results, eight hypotheses on the implementation of formal formative assessment in daily teaching practice in Switzerland were formulated.

9.2 Critique of methodology

The study focussed on the perspective of the teachers who can certainly be seen central in the attempt to bring formal formative assessment methods into regular teaching practice. The perspective of school management representatives; educational politicians; curriculum developers; schoolbook authors; or teacher educators has not been taken into account, however. The perspective of the students was only investigated on a selective level. In that respect, the study does not cover a holistic perspective on the implementation of formative assessment.

The study had a small number of participants: Twenty innovative science teachers from primary and from upper secondary school with many of them having a long-lasting connection to the Fachhochschule Nordwestschweiz. Only two of the primary school teachers worked with grades 1-3. The sample is therefore in no way representative for Swiss science teachers. In particular the results on research question 4 could only be interpreted with caution.

The setting, which was not only designed considering the research questions and the literature, but also taking into account the practical circumstances may have had an influence on the results. The teacher meetings, for example, were held for the teachers from both school levels together for pragmatic reasons. It is not possible to delineate potential effects of this aspect on the results from other effects of the study. A number of similar problems concern the methods of data collection.

Other issues related to the data collection and -analysis: The open setting appeared meaningful to investigate the teachers' perspective on an innovative idea. Many variables could, however, not be controlled, and a question as simple as "did this trial take place in the context of an inquiry-based unit?" or "is that formative assessment?" was, at times, hard to answer. The pragmatic way out was to define theory-based criteria to delineate inquiry-based education from other teaching approaches and formative assessment from other interactions in the classroom. The resulting criteria can only describe the isolated trials on a superficial level and cannot capture deep structures such as the teachers' attitudes towards assessment or their assessment practices throughout a period of time longer than this one unit.

A last consequence of the setting was that the collaboration with the teachers focussed on two themes at the same time: Inquiry-based education and formative assessment. Many of the teachers seemed to perceive these two themes at the same level (formative assessment, inquiry) rather than in a hierarchical relation (formative assessment in the context of inquiry). Particularly in research question 3, it was therefore difficult to delineate between "formative assessment in general" and "formative assessment in the context of inquiry" which would have been a prerequisite for drawing inquiry-specific conclusions. A portion of the hypotheses deduced in chapter 8 might therefore be specific to student-oriented teaching approaches (due to the assessment methods which are particularly suitable in this context and may be less appropriate for

assessing student pre-concepts to plan subsequent steps in teaching) whereas others could be more generally valid. Similarly, a clear distinction between particular methods of formative assessment and formative assessment practices in general was not always possible.

9.3 Implications of the study

The advantage of the small number of participants in the study was the dense picture that could be gained from their trials with different possibilities for triangulation to ensure a certain degree of validity of the results for this group of teachers. Furthermore, the participants were selected so that the two school levels were equally frequent. Within the school levels, the distribution of teaching experience and ages was even. Together with the natural circumstances under which the trial took place, this leads to results with a high expressiveness.

The results provide, on the one hand, first ideas about how to support the uptake of more formal formative assessment in daily teaching practice as claimed nationally and internationally. This could help to plan school books, teacher professional development programmes, and school development programmes as well as the actions at educational ministries. On the other hand, the study also provides ideas on what teacher professional experimentation with formative assessment methods as part of their implementation could look like at the two school levels explored. The two directions of exploration led to a number of hypotheses which will be introduced in the next sub-chapter.

9.4 Prospects

From the two sub-chapters 9.2 and 9.3, need for further research can be inferred: As laid out in 9.2, the study had a small, non-representative sample. The hypotheses that were formulated based on the results of the study should therefore be tested with more teachers and taking into account all school levels. It can be expected that particularly at grades 1-3 might be a different environment for assessment (literacy of the students; students' cognitive abilities; time pressure; role of assessment for selection).

The hypotheses that were deduced from the results of this study conveyed two perspectives: Firstly, ideas about how to support the uptake of more formal formative assessment in daily teaching practice as claimed nationally and internationally.

- H1: Apart from the concept which can also be found in the literature (formative assessment as a means to support student learning and teaching and therefore having a prospective orientation), a number of misconceptions on formative assessment exist amongst teachers from different school levels in Switzerland.
- H2: Not only the teachers from the two school levels in the study but teachers in Switzerland generally have a positive attitude towards using formal formative assessment methods in their inquiry-based science education.
- H3: The aims in terms of student learning which teachers pursue with formal formative assessment methods differ between school levels. Apart from the student learning, the teachers from the two school levels explored also aim at provoking motivational and social effects through formative assessment.
- H4: Formative assessment practices in the context of inquiry-based science education that could realistically be expected from a considerable portion of teachers differ between different school levels.
- H5: In Switzerland, the teachers from both school levels explored perceive several problem areas regarding the implementation of formative assessment: Besides the teacher assessment literacy, the use of lesson time; the availability of resources; and the position of formative assessment within the assessment framework are also considered problematic.

- H6: Measures at different levels might support the uptake of formative assessment activities in the context of Switzerland with its high teacher autonomy and the new curriculum at the compulsory school levels: Classical pre- and in-service professional development; school development projects for the co-construction of knowledge on assessment and for the coordinated introduction of assessment strategies in classes; school books as a source of ideas for formative assessment activities; and a clear communication of the purpose of formative assessment in cantonal and national guidelines.

The second perspective conveyed by the hypotheses are ideas on teacher professional experimentation with formative assessment methods:

- H7: The collaboration of teachers in a study on formative assessment with an open setting as described above, interpreted as an in-service training, can have an effect on the teacher's understanding of and attitude towards formative assessment. The effect could be provoked by different mechanisms: (1) The provision of a theoretical background and inspirational examples on formative assessment; (2) The urge to develop and trial formative assessment activities by asking every teacher to report upon one trial per semester; (3) Encouragement for collaboration between teachers with group discussions where formative assessment activities were exchanged and a dropbox that was used to exchange materials; (4) Encouragement for reflection triggered by the questions in the evaluation form and in the group discussions.
- H8: In the context of formative assessment practices, there are different implementation behaviours. Different variables (school level, gender, teaching experience) appear to have an influence on the implementation behaviour.

10 Literature

- Abd El Khalick, F., Boujaoude, S., Duschl, R. A., Lederman, N. G., Mamlok-Naaman, R., Hofstein, A., Niaz, M., Treagust, D., & Tuan, H. (2004).** Inquiry in science education: International perspectives. *Science Education*, 88(3), 397–419.
- Adamina, M. (2010).** Lernen begleiten, begutachten und beurteilen [Accompanying, surveying and assessing student learning]. In P. Labudde (Ed.), *Fachdidaktik Naturwissenschaft 1. -9. Schuljahr* (pp. 181 - 196). Bern: Haupt Verlag.
- Adams, T. L. (1998).** Alternative assessment in elementary school mathematics. *Childhood Education*, 74, 220–224.
- Ajzen, I. (2005).** Attitudes, personality and behaviour. New York: Open University Press.
- Alkharusi, H. (2011).** Psychometric properties of the teacher assessment literacy questionnaire for pre-service teachers in Oman. *Procedia – Social and Behavioural Sciences*, 29, 1614–1624.
- Allal, L. (2010).** Assessment and the regulation of learning. In P. Peterson, E. Baker, B. McGaw (Eds.), *International encyclopaedia of education*, volume 3 (pp. 248 – 352). Oxford: Elsevier.
- Allal, L. & Lopez, L. M. (2005).** Formative assessment of learning: A review of publications in French. In J. Looney (Ed.), *Formative assessment: Improving learning in secondary classrooms* (pp. 241 – 264). Paris: Organisation for Economic Cooperation and Development.
- American Association for the Advancement of Science (1998).** Blueprints for reform - project 2061; Chapter 8: Assessment. Washington, DC: American Association for the Advancement of Science. Retrieved from <http://www.project2061.org/publications/bfr/online/blpintro.htm> [07.10.2016]
- American Federation of Teachers, National Council on Measurement in Education, & National Education Association (1990).** Standards for teacher competence in educational assessment of students. Washington, DC: National Council on Measurement in Education.
- Andrade, H. (2000).** Using rubrics to promote thinking and learning. *Educational Leadership*, 57(5), 13–18.
- Andrade, H. (2005).** Teaching with rubrics: The good, the bad, and the ugly. *College teaching* 53(1), 27-31.
- Andrade, H. (2010).** Students as the definitive source of formative assessment. In H. Andrade & G.J. Cizek (Ed.), *Handbook of formative assessment*. New York: Routledge.
- Andrade, H. & Du, Y. (2005).** Student perspectives on rubric-referenced assessment. *Practical Assessment, Research & Evaluation*, 10(3), 1–11.
- Andrade, H. & Du, Y. (2007).** Student responses to criteria-referenced self-assessment. *Assessment and Evaluation in Higher Education*, 32, 159–181.
- Andrade, H., Du, Y., & Wang, X. (2008).** Putting rubrics to the test: The effect of a model, criteria generation, and rubric-referenced self-assessment on elementary school students' writing. *Educational Measurement: Issues and Practices*, 27(2), 3–13.
- Andrade, H. & Valtcheva, A. (2009).** Promoting learning and achievement through self-assessment, *Theory into Practice*, 48(1), 12-19.
- Angelo, Th. & Cross, K. P. (1993).** Classroom assessment techniques. A handbook for college teachers. San Francisco: Jossey-Bass.
- ARG (Assessment Reform Group) (2002).** Assessment for learning: 10 Principles. London: ARG. Retrieved from <http://www.assessment-reform-group.org> [07.10.2016]
- Arter, J. A. & McTighe, J. (2001).** Scoring rubrics in the classroom: Using performance criteria for assessing and improving student performance. Thousand Oaks: Corwin Press.
- Artigue, M. & Baptist, P. (2012).** Inquiry in mathematics education. Background resources for implementing inquiry in science and mathematics at school. Paris: Université Paris Diderot. Retrieved from <http://www.fibonacci-project.eu/> [07.10.2016]
- Ashton, P. & Webb, R. (1986).** Making a difference: Teachers' sense of efficacy and student achievement. New York: Longman.
- Ayala, C. C., Shavelson, R. J., Ruiz-Primo, M. A., Brandon, P. R., Yin, Y., Furtak, E. M., Young, D. B., & Tomita, M. K. (2008).** From formal embedded assessments to reflective lessons: The development of formative assessment studies. *Applied Measurement in Education*, 21(4), 315–334.

- Bailey, R. & Garner, M. (2010).** Is the feedback in higher education assessment worth the paper it is written on? Teachers' reflections on their practices. *Teaching in higher education* 15(2), 187-198.
- Bandura, A. (1977).** Self-efficacy: Toward a unifying theory of behavioural change. *Psychological Review*, 84, 191-215.
- Bandura, A. (1982).** Self-efficacy mechanism in human agency. *American Psychologist*, 37, 122-147.
- Bandura, A. (1997).** Self-efficacy: The exercise of control. New York: W. H. Freeman and Company.
- Barron, B. & Darling-Hammond, L. (2008).** Teaching for meaningful learning: A review of research on inquiry-based and cooperative learning. In L. Darling-Hammond, B. Barron, P. D. Pearson, A. H. Schoenfeld, E. K. Stage, T. D. Zimmermann, G. N. Cervetti, & J. Tilson (Eds.), *Powerful Learning. What we know about teaching for understanding* (pp. 11-70). San Francisco: Jossey-Bass.
Retrieved from <http://www.edutopia.org/pdfs/edutopia-teaching-for-meaningful-learning.pdf> [07.10.2016]
- Bell, B. & Cowie, B. (2001).** The characteristics of formative assessment in science education. *Science Education*, 85(5), 536-553.
- Bell, T., Urhahn, D., Schanze, S., & Ploetzner, R. (2010).** Collaborative inquiry learning: Models, tools, and challenges. *International Journal of Science Education*, 32(3), 349-377.
- Bennett R. E. (2011).** Formative assessment: a critical review. *Assessment in Education: Principles, Policy & Practice*, 18(1), 5-25.
- Bharuthram, S. (2015).** Lecturers' perceptions: The value of assessment rubrics for informing teaching practice and curriculum review and development. *Africa education review*, 12(3), 415 - 428.
- Bhattacharyya, S., Volk, T., & Lumpe, A. (2009).** The influence of an extensive inquiry-based field experience on pre-service elementary student teachers' science teaching beliefs. *Journal of Science Teacher Education*, 20, 199 - 218.
- Birenbaum, M., Breuer, K., Cascallar, E., Dochy, F., Dori, Y., Ridgway, J., Wiesemes, R. (Ed.), & Nickmans, G. (Ed.) (2006).** A learning integrated assessment system. *Educational Research Review*, 1, 61-67.
- Black, P. (1993).** Formative and summative assessments by teachers. *Studies in Science Education*, 21, 49-97.
- Black, P. & Atkin, J.M. (1996).** Changing the subject: Innovations in science, mathematics and technology education, London: Routledge for OECD.
- Black, P. & Harrison, Ch. (2001).** Self- and peer-assessment and taking responsibility: The science student's role in formative assessment. *School science review* 83(302), 43-49.
- Black, P. & Harrison, Ch. (2004).** Science inside the black box. London: GL Assessment.
- Black, P., Harrison, Ch., Lee, C., Marshall, B., & Wiliam, D. (2003).** Assessment for learning: Putting it into practice. London: Open University Press.
- Black, P., Harrison, Ch., Lee, C., Marshall, B., & Wiliam, D. (2004).** Working inside the black box: Assessment for learning in the classroom. Phi Delta Kappan.
- Black, P. & Wiliam, D. (1998).** Assessment and classroom learning. *Assessment in Education: Principles, Policy and Practice*. 5(1), 7-73.
- Black, P. & Wiliam, D. (2009).** Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability*, 21(1), 5-13.
- Blömeke, S., Risse, J., Müller, C., Eichler, D. & Schulz, W. (2006).** Analyse der Qualität von Aufgaben aus didaktischer und fachlicher Sicht. Ein allgemeines Modell und seine exemplarische Umsetzung im Unterrichtsfach Mathematik [Analysis of the quality of tasks from an educational and content-specific perspective. A general model and its implementation in mathematics education]. *Unterrichtswissenschaft* 34(4), 330-357.
- Bloom, B. S. (1969).** Some theoretical issues relating to educational evaluation. In R. W. Tyler (Ed.), *National Society for the Study of Education Yearbook: 68* (2). Educational evaluation: New roles, new means (pp. 26-50). Chicago: University of Chicago Press.
- Binkley, M., Erstad, O., Herman, J. L., Raizen, S., Ripley, M., Miller-Ricci, M., & Rumble, M. (2012).** Defining twenty-first century skills. In P. E. Griffin, B. McGaw, & E. Care (Eds.), *Assessment and teaching of 21st century skills* (pp. 17-66). Dordrecht, New York: Springer.

- Bose, J. & Rengel, Z. (2009).** A model formative assessment strategy to promote student-centred self-regulated learning in higher education. *US-China Education Review*, 6(12), 29–35.
- Börlin, J. (2012).** Das Experiment als Lerngelegenheit. Vom interkulturellen Vergleich des Physikunterrichts zu Merkmalen seiner Qualität [The experiment as an opportunity to learn. From the intercultural comparison of physics education to characteristics of its quality]. Berlin: Logos Verlag.
- Börlin, J. & Labudde, P. (2014).** Practical work in physics instruction: An opportunity to learn? In H. E. Fischer, P. Labudde, K. Neumann & J. Viiri (Eds.), *Quality of instruction in physics* (pp. 111 – 128). Münster: Waxmann.
- Boud, D. (1990).** Assessment and the promotion of academic values. *Studies in Higher Education*, 15(1), 101-111.
- Burke, K. (2006).** From standards to rubrics in 6 steps. Heatherton, Victoria: Hawker Brownlow Education.
- Breidenstein, G., Meier, M., & Zaborowski, K. U. (2012).** Die Ethnographie schulischer Leistungsbewertung. Ein Beispiel für qualitative Unterrichtsforschung [Ethnography of assessment of student achievement in the context of education. An example of qualitative research in education]. In F. Ackermann, T. Ley, C. Machold, M. Schrödter (Eds.), *Qualitatives Forschen in der Erziehungswissenschaft* (pp. 157 - 175). Wiesbaden: Springer VS Verlag.
- Brígido, M., Borrachero, A. B., Bermejo, M. L., & Mellado, V. (2013).** Prospective primary teachers' self-efficacy and emotions in science teaching. *European Journal of Teacher Education*, 36(2), 200-217.
- Boekaerts, M. & Corno, L. (2005).** Self-regulation in the classroom: A perspective on assessment and intervention. *Applied Psychology: An International Review*, 54(2), 199–231.
- Brookhard, M., Moss, C. M., & Long, B. A. (2010).** Teacher inquiry into formative assessment practices in remedial reading classrooms. *Assessment in Education*, 17(1), 41 – 58.
- Brookhart, S. M. (2011).** Educational assessment knowledge and skills for teachers. *Educational Measurement: Issues and Practice*, 30(1), 3–12.
- Brown, G. T. L. (2004).** Teachers' conceptions of assessment: Implications for policy and professional development. *Assessment in Education: Principles, Policy & Practice*, 11(3), 301-318.
- Brown, G. T. L. (2008).** *Conceptions of assessment: Understanding what assessment means to teachers and students.* New York: Nova Science Publishers.
- Brown, G. T. L., Harris, L. R., & Harnett, J. (2012).** Teacher beliefs about feedback within an assessment for learning environment: Endorsement of improved learning over student well-being. *Teaching and teacher education*, 28, 968 – 978.
- Bruno, I. & Santos, L. (2010).** Written comments as a form of feedback. *Studies in educational evaluation* 36, 111-120.
- Bürgermeister, A., Klimczak, M., Klieme, E., Rakoczy, K., Blum, W., Leiss, D., Harks, B., & Besser, M. (2011).** Leistungsbeurteilung im Mathematikunterricht. Eine Darstellung des Projekts "Nutzung und Auswirkungen der Kompetenzmessung in mathematischen Lehr-Lernprozessen" [Assessment of student achievement in mathematics education. A synopsis of the project „Usage and effects of the measurement of competences in mathematical learning and teaching]. *Schulpädagogik - heute*, 2(3), 1-18.
- Bybee, R. (1997).** *Achieving scientific literacy: From purposes to practices.* Portsmouth: Heilmann.
- Bybee, R. (2000).** Teaching science as inquiry. In J. Minstrell & E. van Zee (Eds.), *Inquiring into inquiry learning and teaching in science* (pp. 20-46). Washington DC: American Association for the Advancement of Science.
Retrieved from <https://www.aaas.org> [07.10.2016]
- Calderhead, J. (1996).** Teachers' beliefs and knowledge. In D. C. Berliner & R. C. Calfee (Eds), *Handbook of educational psychology* (709-725). New York: Simon & Schuster Macmillan.
- Chudowsky, N. & Pellegrino, J. W. (2003).** Large-scale assessments that support learning: What will it take? *Theory into Practice*, 42(1), 75–83.
- Cizek, G. (2009).** Reliability and validity of information about student achievement: Comparing the contexts of large-scale and classroom testing. *Theory into practice*, 48(1), 63-71.
- Cizek, G. (2010).** An introduction to formative assessment. In H. Andrade & G.J. Cizek (Ed.), *Handbook of formative assessment* (pp. 3-17). New York: Routledge.
- Clark, C. & Peterson, P. (1986).** Teachers' thought processes: In M. Wittrock (Ed.), *Handbook of research on teaching* (pp. 255-296). New York: MacMillan.

- Clark, I. (2012).** Formative assessment: Assessment is for self-regulated learning. *Educational Psychology Review*, 24, 205 – 249.
- Clarke, D. & Hollingsworth, H. (2002).** Elaborating a model of teacher professional growth. *Teaching and teacher education*, 18, 947 – 967.
- Cohen, J. (1988).** *Statistical power analysis for the behavioural sciences*. Hillsdale: Lawrence Erlbaum Associates.
- Cone, N. (2009).** Pre-service elementary teachers' self-efficacy beliefs about equitable science teaching: Does service learning make a difference? *Journal of Science Teacher Education*, 21(2), 25–34.
- Cole, D.A. (1991).** Change in self-perceived competence as a function of peer and teacher evaluation. *Developmental Psychology*, 27, 682-688.
- Connors, R.J. & Lunsford, A.A. (1993).** Teachers' rhetorical comments on student papers. *College Composition and Communication*, 44, 200–223.
- Cowie, B. & Bell, B. (1999).** A model for formative assessment. *Assessment in Education* 6(1), 101-116.
- Crooks, T. J. (1988).** The impact of classroom evaluation practices on students. *Review of Educational Research*, 58, 438-481.
- Cuevas, P., Lee, O., Hart, J., & Deaktor, R. (2005).** Improving science inquiry with elementary students of diverse backgrounds. *Journal of Research in Science Teaching*, 42(3), 337–357.
- Czerniak, C.M. (1990).** A study of self-efficacy, anxiety, and science knowledge in pre-service elementary teachers. A paper presented at the annual meeting of the National Association of Research in Science Teaching, Atlanta, GA.
- D-EDK Deutschschweizer Erziehungsdirektoren-Konferenz (2014).** *Lehrplan 21 [Curriculum 21]*. Luzern: D-EDK.
Retrieved from <http://vorlage.lehrplan.ch/downloads.php> [07.10.2016]
- De Jong, T. & Njoo, M. (1992).** Learning and instruction with computer simulations: Learning processes involved. In E. de Corte, M. Linn, H. Mandl, & L. Verschaffel (Eds.), *Computer-based learning environments and problem solving* (pp. 411–427). Berlin: Springer.
- De Jong, T. & van Joolingen, W. R. (1998).** Scientific discovery learning with computer simulations of conceptual domains. *Review of Educational Research*, 68(2), 179–201.
- Dempster, F. N. (1991).** Synthesis of research on reviews and tests. *Educational leadership*, 48(7), 71-76.
- Dempster, F. N. (1992).** Using tests to promote learning: A neglected classroom resource. *Journal of Research and Development in Education*, 25(4), 213 – 217.
- Desimone, L. (2009).** Improving impact studies of teachers' professional development: Toward better conceptualizations and measures. *Educational Researcher*, 38, 181–199.
- Dixon, H. (1999).** *The effect of policy on practice: An analysis of teachers' perceptions of school based assessment practice*. Albany, NZ: Massey University.
- Dochy, F. & Moerkerke, G. (1997).** The present, the past and the future of achievement testing and performance assessment. *International Journal of Educational Research*, 27, 415-432.
- Dochy, F., Segers, M., & Sluijsmans, D. (1999).** The use of self-peer and co-assessment in higher education: A review. *Studies in Higher Education*, 24(3), 331-350.
- Dolin, J. (2012).** *Assess inquiry in science, technology and mathematics education: ASSIST-ME proposal*. Copenhagen: University of Copenhagen.
Retrieved from <http://assistme.ku.dk> [07.10.2016]
- Dolin, J. & Evans, R. H. (2013).** The contribution of formative assessment and self-efficacy to inquiry learning. In M.H. Hoveid & P. Gray (Eds.), *Inquiry in science education and science teacher education* (pp. 125-145). Trondheim: Akademika Publishing.
- Dreyer, H. P. (2015).** Eine MINT-Initiative für das Gymnasium [A STEM initiative for the Gymnasium]. *Gymnasium Helveticum*, 1, 11-13.
- Duffrin, N., Dawes, W., Hanson, D., Miyazaki, J., & Wolfskill, T. (1998).** Transforming large introductory classes into active learning environments. *Journal of Educational Technology Systems*, 27, 169–78.
- Duschl, R. A., (2003).** Assessment of inquiry. In J.M. Atkin & J.E. Coffey (Eds.), *Everyday assessment in the science classroom* (pp. 41 – 59). Arlington, Virginia: NSTA press.

- Dusenbury, L., Brannigan, R., Falco, M., & Hansen, W. B. (2003).** A review of research on fidelity of implementation: Implications for drug abuse prevention in school settings. *Health Education Research*, 18, 237-256.
- Dzelili, A. (2009).** Die grosse Frage im Hintergrund: Wozu ist die Schule da? [The big underlying question: What is the purpose of school?] In D. Fischer, A. Strittmatter, U. Vögeli-Mantovani (Eds.), *Noten, was denn sonst? Leistungsbeurteilung und -bewertung* (pp. 41-46). Zürich: Verlag LCH.
- EDK Schweizerische Konferenz der kantonalen Erziehungsdirektoren (1994).** Rahmenlehrplan für die Maturitätsschulen [Curricular framework for upper secondary schools]. Bern: Schweizerische Konferenz der kantonalen Erziehungsdirektoren.
- EDK Schweizerische Konferenz der kantonalen Erziehungsdirektoren (2011).** Grundkompetenzen für die Naturwissenschaften. Nationale Bildungsstandards [Basic competences for science education. National educational standards]. Bern: Schweizerische Konferenz der kantonalen Erziehungsdirektoren.
- Engel, N. (2008).** Förderdiagnostik in der Alphabetisierung. Eine empirische Untersuchung zur Schreibprozessdiagnose in Alphabetisierungskursen Niedersachsens [Supportive diagnostics in the alphabetisation. An empirical study on the diagnosis of writing processes in courses on alphabetisation in Niedersachsen]. Stuttgart: ibidem.
- Enochs, L. & Riggs, I. (1990).** Further development of an elementary science teaching efficacy belief instrument: A preservice elementary scale. *School Science and Mathematics*, 90, 694-706.
- Euler, M. (2011).** PRIMAS survey report on inquiry-based learning and teaching in Europe. Kiel: IPN Kiel. Retrieved from <http://www.primas-project.eu> [07.10.2016]
- European Commission (2004).** Increasing human resources for science and technology in Europe: Report of the high level group on human resources for science and technology in Europe. Luxembourg: Office for Official Publications of the European Communities.
- Evans, K. A. (2001).** Rethinking self-assessment as a tool for response. *Teaching English in the Two-Year College*, 28, 293-301.
- Falchikov, M. (1991).** Group process analysis. In S. Brown & P. Dove (Eds.), *Self- and peer-assessment* (pp.15-27). Birmingham, Standing conference on educational development, paper 63.
- Falchikov, N. & Boud, D. (1989).** Student self-assessment in higher education: A meta-analysis. *Review of Educational Research*, 59, 395-430.
- Ferguson, P. (2009).** Student perceptions of quality feedback in teacher education. *Assessment and Evaluation in Higher Education*, 36, 51-62.
- Firestone, W. A., Mayrowetz, D., & Fairman, J. (1998).** Performance-based assessment and instructional change: The effects of testing in Maine and Maryland. *Educational evaluation and policy analysis*, 20(2), 95-113.
- Fischer, D. (2009).** Keine Noten - keine Beurteilung? [No grades – no assessment?] In D. Fischer, A. Strittmatter, U. Vögeli-Mantovani (Eds.), *Noten, was denn sonst? Leistungsbeurteilung und -bewertung* (pp. 25-26). Zürich: Verlag LCH.
- Fischer, H.-E. & Draxler, D. (2001).** Aufgaben und naturwissenschaftlicher Unterricht [Tasks and science education]. *MNU journal*, 54(7), 388-393.
- Frey, K. & Frey-Eiling, A. (2004).** Allgemeine Didaktik [Educational sciences]. Zürich: Verlag der Fachvereine Zürich.
- Friedler, Y., Nachmias, R., & Linn, M. C. (1990).** Learning scientific reasoning skills in micro-computer-based laboratories. *Journal of Research in Science Teaching*, 27(2), 173-191.
- Fulton, K. & Britton, T. (2010).** STEM teachers in professional learning communities: A knowledge synthesis. Washington, DC: National Commission on Teaching and America's Future.
- Furtak, E. M. & Ruiz-Primo, M. A. (2008).** Making students' thinking explicit in writing and discussion. An analysis of formative assessment prompts. *Science Education* 92(5), 799-824.
- Furtak, E. M., Ruiz-Primo, M. A., Shemwell, J. T., Ayala, C., Brandon, P. R., Shavelson, R. J., & Yin, Y. (2008).** On the fidelity of implementing embedded formative assessments and its relation to student learning. *Applied measurement in education*, 21(4), 360-389.
- Furtak, E. M., Seidel, T., Iverson, H., & Briggs, D. C. (2012).** Experimental and quasi-experimental studies of inquiry-based science teaching: A meta-analysis. *Review of Educational Research*, 82(3), 300-329.

- Gardner, J., Harlen, W., Hayward, L., Stobart, G., & Montgomery, M. (2010).** Developing teacher assessment. Maidenhead: Open University Press.
- Gijlers, H. & de Jong, T. (2005).** The relation between prior knowledge and students' collaborative discovery learning processes. *Journal of Research in Science Teaching*, 42(3), 264–282.
- Gitomer, D. H. & Duschl, R. A. (2007).** Establishing multilevel coherence in assessment. In P.A. Moss (Ed.), Evidence and decision making. The 106th yearbook of the National Society for the Study of Education, Part I (pp. 288–320). Chicago, IL: National Society for the Study of Education.
- Glover, C. & Brown, E. (2006).** Written feedback for students: Too much, too detailed or too incomprehensible to be effective? *Bioscience Education*, 7(1), 1-16.
- Goodrich, H. (1996).** Student self-assessment: At the intersection of metacognition and authentic assessment. Cambridge, MA: Harvard University.
- Gregory, K., Cameron, C., & Davies, A. (2000).** Self-assessment and goal-setting. Merville, Canada: Connection.
- Gresham, F. M. (1989).** Assessment of treatment integrity in school consultation and prereferral intervention. *School Psychology Review*, 18, 37–50.
- Grob, U. & Maag Merki, K. (2001).** Überfachliche Kompetenzen. Theoretische Grundlegung und empirische Erprobung eines Indikatorensystems [Cross-curricular competences. Theoretical background and empirical validation of a system of indicators]. Bern: Peter Lang.
- Grotlüschen, A. & Bonna, F. (2008).** German-language review in Teaching, learning and assessment for adults: Improving foundation skills. Paris: OECD Publishing.
Retrieved from <https://www.oecd.org/edu/ceri/40046802.pdf> [07.10.2016]
- Hanrahan, S. J. & Isaacs, G. (2001).** Assessing self- and peer-assessment: The students' views. *Higher Education Research and Development*, 20, 53–70.
- Hargreaves, E. & McCallum, B. (1998).** Written feedback to children from teachers. ESRC project 'Teaching, Assessment and Feedback Strategies', Project Paper No. 7. London: University of London.
- Harlen, W. (2007).** Holding up a mirror to classroom practice. *Primary Science Review*, 100, 29–31.
- Harlen, W. (2009).** Teaching and learning science for a better future. *School Science Review*, 90(333), 33–41.
- Harlen, W. (2013).** Assessment & inquiry-based science education: Issues in policy and practice. Trieste: Global Network of science Academies (IAP) science Education Programme (SEP).
Retrieved from <http://www.interacademies.net/File.aspx?id=21245> [07.10.2016]
- HarmoS: Konsortium HarmoS Naturwissenschaften (2008).** HarmoS Naturwissenschaften+. Kompetenzmodell und Vorschläge für Bildungsstandards. Wissenschaftlicher Schlussbericht [HarmoS science education+. Model of competences and suggestions for educational standards. Final scientific report]. Bern: Konsortium HarmoS Naturwissenschaften.
- Harms, U., Mayer, R. E., Hammann, M., Bayrhuber, H., & Kattmann, U. (2004).** Kerncurriculum und Standards für den Biologieunterricht in der gymnasialen Oberstufe [Curriculum and standards for biology education at the upper secondary school level]. In H.-E. Tenorth (Ed.), *Kerncurriculum Oberstufe II. Biologie, Chemie, Physik, Geschichte, Politik* (pp. 22–84). Weinheim: Beltz.
- Harrington, T. (1995).** Assessment of abilities. Greensboro, NC: ERIC Clearinghouse on Counselling and Student Services.
- Hart, D. (1999).** Opening assessment to our students. *Social Education*, 63, 343–345.
- Hartig, J., Klieme, E., & Leutner, D. (2008).** Assessment of competencies in educational contexts. Göttingen: Hogrefe Publishing GmbH.
- Harvey, L. & Knight, P. T. (1996).** Transforming higher education. Buckingham, Society for Research into Higher Education and the Open University Press.
- Hattie, J. (2009).** Visible learning. A synthesis of over 800 meta-analyses relating to achievement. London & New York: Routledge.
- Hattie, J. & Timperley, H. (2007).** The power of feedback. *Review of Educational Research* 77(1), 81-112.
- Hechter, R. (2011).** Changes in preservice elementary teachers' personal science teaching efficacy and science teaching outcome expectancies: The influence of context. *Journal of Science Teacher Education*, 22, 187–208.

- Heritage, M. (2010).** *Formative assessment: Making it happen in the classroom.* Thousand Oaks, California: Corwin Press.
- Herman, J. L., Osmundson, E. & Silver, D. (2010).** Capturing quality in formative assessment practice: Measurement challenges, CRESST Report 770. Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing (CRESST).
- Hernández, R. (2012).** Does continuous assessment in higher education support student learning? *Higher Education*, 64, 489–502.
- Hill, M. F. (2000).** Remapping the assessment landscape: Primary teachers reconstructing assessment in self-managing schools. Hamilton, NZ: University of Waikato.
- Hinrichsen, J. & Jarrett, D. (1999).** *Science inquiry for the classroom: A literature review.* Portland: Northwest Regional Educational Laboratory.
- Hofstein, A., Navon, O., Kipnis, M., & Mamlok-Naaman, R. (2005).** Developing students' ability to ask more and better questions resulting from inquiry-type chemistry laboratories. *Journal of Research in Science Teaching*, 42(7), 791-806.
- Hondrich, A. L., Hertel, S., Adl-Amini, K. & Klieme, E. (2016).** Implementing curriculum-embedded formative assessment in primary school science classrooms. *Assessment in Education: Principles, Policy & Practice*, 23(3), 353-376.
- Hord, S. (1997).** *Professional learning communities: Communities of continuous inquiry and improvement.* Austin, TX: Southwest Educational Development Laboratory.
- Husfeld, V. (2009).** Aus der Praxis der Leistungsbeurteilung [On the practice of summative assessment]. In D. Fischer, A. Strittmatter, U. Vögeli-Mantovani (Eds.), *Noten, was denn sonst? Leistungsbeurteilung und -bewertung* (pp. 33-40). Zürich: Verlag LCH.
- Hyland, F. (1998).** The impact of teacher written feedback on individual writers. *Journal of Second Language Writing*, 7(3), 255-286.
- Irving, S., Harris, L., & Peterson, E. (2011).** 'One assessment doesn't serve all the purposes' or does it? New Zealand teachers describe assessment and feedback. *Asia Pacific Education Review*, 12(3), 413 - 426.
- Jonsson, A. (2014).** Rubrics as a way of providing transparency in assessment. *Assessment and Evaluation in Higher Education*, 39(7), 840 - 852.
- Jorde, D., Olsen Moberg, A., Rønnebeck, S., & Stadler, M. (2012).** *Work package 2: Final report.* Trondheim: University of Trondheim.
- Jundt, W. (2013).** Unpassendes zur Beurteilung [Inappropriate things on assessment]. *profil Magazin für das Lehren und Lernen*, 1, 10-11.
- Keeley, P. (2008).** *Science formative assessment. 75 practical strategies for linking assessment, instruction, and learning.* California: Corwin Press.
- Kessler, J. H. & Galvan, P. M. (2007).** *Inquiry in action: Investigating matter through inquiry.* Washington DC: American Chemical Society.
Retrieved from <http://www.inquiry-inaction.org/download/> [07.10.2016]
- Kind, P. M. & Kind, V. (2007).** Creativity in science education: Perspectives and challenges for developing school science. *Studies in Science Education*, 43(1), 1-37.
- Kluger, A. N. & DeNisi, A. (1996).** The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological bulletin*, 119(2), 254 - 284.
- Kölller, O. (2005).** Formative assessment in classrooms: A review of empirical German literature. In J. Looney (Ed.), *Formative assessment: improving learning in secondary classrooms* (pp. 265 - 279). Paris: OECD Publishing.
- Kronig, W. (2009).** Schulnoten - Glasperlen des Bildungssystems [Grades - Glass beads of the educational system]. In D. Fischer, A. Strittmatter, U. Vögeli-Mantovani (Eds.), *Noten, was denn sonst? Leistungsbeurteilung und -bewertung* (pp. 27-32). Zürich: Verlag LCH.
- Kulhavy, R. W. (1977).** Feedback in written instruction. *Review of Educational Research*, 47(1), 211-232.
- Kwan, K. & Leung, R. (1996).** Tutor versus peer group assessment of student performance in a stimulation training exercises. *Assessment and Evaluation in Higher Education*, 21(3), 205-214.

- Labudde, P. (2000).** Konstruktivismus im Physikunterricht der Sekundarstufe II [Constructivism in upper secondary physics education]. Bern / Stuttgart / Wien: Haupt Verlag.
- Labudde, P. (2007).** How to develop, implement and assess standards in science education? 12 challenges from a Swiss perspective. In D. Washington, P. Nentwig & S. Schanze (Eds.), *Making it comparable: Standards in science education* (pp. 277 – 301). Münster: Waxmann.
- Labudde, P., Nidegger, Ch., Adamina, M., & Gingins, F. (2007).** The development, validation, and implementation of standards in science education: Chances and difficulties in the Swiss project HarmoS. In D. Washington, P. Nentwig & S. Schanze (Eds.), *Making it comparable: Standards in science education* (pp. 235 – 259). Münster: Waxmann.
- Landis, J. R. & Koch, G. G. (1977).** The measurement of observer agreement for categorical data. *Biometrics*, 33, 159–174.
- Leahy, S., Lyon, Ch., Thompson, M., & William, D. (2005).** Classroom assessment: Minute by minute, day by day. *Assessment to promote learning*, 63(3), 19-24.
- Leki, I. (2006).** “You cannot ignore”: L2 graduate students’ response to discipline-based written feedback. In K. Hyland, F. Hyland (Eds.), *Feedback in second language writing* (pp. 266–286). Cambridge: Cambridge University Press.
- Lin, S. S. J., Liu, E. Z. F., & Yuan, S. M. (2001).** Web-based peer assessment: Feedback for students with various thinking styles. *Journal of Computer Assisted Learning*, 17, 420-432.
- Lindsay, C. & Clarke, S. (2001).** Enhancing primary science through self- and paired-assessment. *Primary science Review*, 68, 15–18.
- Löhner, S., van Joolingen, W. R., Savelsbergh, E. R., & van Hout-Wolters, B. (2005).** Students’ reasoning during modelling in an inquiry learning environment. *Computers in Human Behavior*, 21, 441–461.
- Looney, J. W. (2011).** Integrating formative and summative assessment: Progress toward a seamless system? OECD Education Working Papers No. 58. Paris: OECD Publishing.
- Looney, J., Laneve, C., & Moscato, M. T. (2005).** Italy: A system in transition. In *Organization of Economic Co-operation and Development, Formative assessment: Improving learning in secondary classrooms* (pp. 163–175). Paris: OECD Publishing.
- Luft, J. A. (1999).** Rubrics: Design and use in science teacher education. *Journal of Science Teacher Education*, 10(2), 107–121.
- Lunetta, V. N. (1998).** The school science laboratory: Historical perspectives and context for contemporary teaching. In B. Fraser & K. Tobin (Eds.), *International handbook of science education* (pp. 249–264). Dordrecht, The Netherlands: Kluwer.
- Maier, U. (2011).** Formative Leistungsdiagnostik in der Sekundarstufe I – Befunde einer quantitativen Lehrerbefragung zu Nutzen und Korrelaten verschiedener Typen formativer Diagnosemethoden in Gymnasien [Formative assessment at lower secondary school – Quantitative results of a teacher survey on usage and correlates of different types of formative assessment methods at the Gymnasium]. *Empirische Pädagogik*, 25 (1), 25 – 46.
- Maier, U. (2015).** Leistungsdiagnostik in Schule und Unterricht. Schülerleistungen messen, bewerten und fördern [Formative assessment in education. Measuring, assessing and supporting student achievement]. Bad Heilbrunn: Verlag Julius Klinkhardt.
- Mansell, W., James, M. & the Assessment Reform Group. (2009).** *Assessment in schools. Fit for purpose?* A commentary by the Teaching and Learning Research Programme. London: TLRP.
- Mann, H. B. & Whitney, D. R. (1947).** On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, 18(1), 50–60.
- Marshall, B. & Drummond, M. J. (2006).** How teachers engage with assessment for learning: Lessons from the classroom. *Research Papers in Education*, 21(2), 133-149.
- Mayring, Ph. (1994).** *Qualitative Inhaltsanalyse. Grundlagen und Techniken* [Qualitative content analysis. Theoretical background and techniques]. Weinheim: Deutscher Studien Verlag.
- Mayring, Ph. (2010).** *Qualitative Inhaltsanalyse. Grundlagen und Techniken*. [Qualitative content analysis. Theoretical background and techniques]. Weinheim: Beltz.
- McLoughlin, E., Finlayson, O., & van Kampen, P. (2012).** *SAILS – Report on mapping the development of key skills and competencies onto skills developed in IBSE: WP 1 – Deliverable 1.1*. Dublin: Dublin City University.

- Retrieved from <http://www.sails-project.eu/sites/default/files/outcomes/d1-1.pdf> [07.10.2016]
- Mertler, C. A. (2009).** Teachers' assessment knowledge and their perceptions of the impact of classroom assessment professional development. *Improving schools*, 12(2), 101 – 113.
- Millar, R., Tiberghien, A. & Le Maréchal, J.-F. (2002).** Varieties of labwork: A way of profiling labwork tasks. In D. Psillos & H. Niedderer (Eds.), *Teaching and Learning in the Science Laboratory* (pp. 9–20). Dordrecht: Kluwer.
- Mintzes, J. J., Marcum, B., Messerschmidt-Yates, Ch., & Mark, A. (2013).** Enhancing self-efficacy in elementary science teaching with professional learning communities. *Journal in Science Teacher Education*, 24, 1201 – 1218.
- Moni, R. W. & Moni, K. B. (2008).** Student perceptions and use of an assessment rubric for a group concept map in physiology. *Advances in Physiology Education*, 32(1), 47– 54.
- Morrison, J. A. & Lederman, N. G. (2003).** Science teachers' diagnosis and understanding of students' pre-conceptions. *Science Education*, 87, 849–867.
- Moskal, B. M. (2003).** Recommendations for developing classroom performance assessments and scoring rubrics. *Practical Assessment, Research & Evaluation*, 8 (14).
Retrieved from <http://pareonline.net/getvn.asp?V=8&n=14> [07.10.2016]
- Natriello, G. (1987).** The impact of evaluation processes on students. *Educational Psychologist*, 22, 155–175.
- NGSS Lead States (2013).** *Next Generation Science Standards: For States, By States*. Washington, DC: The National Academies Press.
- Nicol, D. & Macfarlane-Dick, D. (2006).** Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education*, 31, 199–218.
- Ni, Y. (1997).** Performance-based assessment: Problems and design strategies. *Education Journal*, 25(2), 137–57.
- NRC (National Research Council) (1996).** *National science education standards*. Washington, D.C.: The National Academies Press.
- NRC (National Research Council) (2011).** *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. Washington DC: National Academic Press.
- Nunes, C. (2004).** *A avaliação como regulação do processo de ensino aprendizagem da Matemática [Assessment and regulation of the teaching in mathematics education]*. Lisbon: Lisbon University.
- OECD (Organization for Economic Co-operation and Development) (2005a).** *Formative assessment. Improving learning in secondary classrooms*. Paris, France: OECD Publishing.
- OECD (Organization for Economic Co-operation and Development) (2005b).** *The definition and selection of key competences. Executive summary*. Paris, France: OECD Publishing.
- OECD (Organization for Economic Co-operation and Development) (2013).** *Synergies for better learning: An international perspective on evaluation and assessment. OECD Reviews of Evaluation and Assessment in Education*. OECD Publishing, Paris.
- Pajares, M. F. (1992).** Teachers' beliefs and educational research: Cleaning up a messy construct. *Review of Educational Research*, 62, 307-332.
- Palmer, D. (2006).** Sources of self-efficacy in a science methods course for primary teacher education students. *Research in Science Education*, 36, 337–353.
- Panadero, E. & Alonso-Tapia, J. (2013).** Self-assessment: Theoretical and practical connotations. When it happens, how is it acquired and what to do to develop it in our students. *Electronic Journal of Research in Education & Psychology*, 11(2), 551 – 576.
- Panadero, E. & Jonsson, A. (2013).** The use of scoring rubrics for formative assessment purposes revisited: A review. *Educational Research Review*, 9(1), 129–144.
- Paris, S. G. & Paris, A. H. (2001).** Classroom applications of research on self-regulated learning. *Educational Psychologist*, 36, 89–101.
- Pedder, D. (2006).** Organizational conditions that foster successful classroom promotion of learning how to learn. *Research Papers in Education*, 21(2), 171–200.
- Perrenoud, Ph. (1991).** Pour une approche pragmatique de l'évaluation formative [For a pragmatic approach to formative assessment]. *Mesure et évaluation en éducation*, 13(4), 49-81.

- Perrenoud, Ph. (1998).** From formative evaluation to a controlled regulation of learning processes. Towards a wider conceptual field. *Assessment in education: Principles, Policy and Practice*, 5(1), 85 – 102.
- Pellegrino, J. W. & Hilton, M. L. (2012).** Education for life and work: Developing transferable knowledge and skills in the 21st century. Washington, D.C: The National Academies Press.
- Pintrich, P. (2000).** The role of goal orientation in self-regulated learning. In M. Boekaerts, P. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 452–502). San Diego, CA: Academic Press.
- Pintrich, P. R. & Zusho, A. (2002).** The development of academic self-regulation: the role of cognitive and motivational factors. In A. Wigfield, & J. S. Eccles (Eds.), *Development of achievement motivation* (pp. 249–284). San Diego: Academic Press.
- Popham, W. J. (2008).** Classroom assessment: What teachers need to know. Boston: Prentice Hill.
- Postholm, M. B. (2012).** Teachers' professional development: A theoretical review. *Educational Research*, 54, 405–429.
- Priemer, B. (2011).** Was ist das Offene beim offenen Experimentieren? [What is open in open experimentation?] *Zeitschrift für Didaktik der Naturwissenschaften*, 17, 315 – 337.
- Ramaprasad, A. (1983).** On the definition of feedback. *Behavioural science*, 28(1), 4-13.
- Ramey-Gassert, L., & Shroyer, M. G. (1986).** Enhancing science teaching self-efficacy in preservice elementary teachers. *Journal of Elementary Science Education*, 4(1), 26–34.
- Ross, J. A., Hogaboam-Gray, A., & Rolheiser, C. (2002).** Student self-evaluation in grade 5–6 mathematics effects on problem-solving achievement. *Educational Assessment*, 8, 43–59.
- Rothenbacher, M. (2010).** Beurteilen im Mathematikunterricht mit dem Zahlenbuch: Begriffs-Klärungen, -Verständnis, -Grundlagen [Assessment in mathematics education with the Zahlenbuch: Clarification of terms and theoretical backgrounds]. Retrieved from http://www.zahlenbu.ch/cms/media/archive3/kursunterlagen_zahlenbuch/WB_Kurs_Beurteilung_MATH_Grundlagen_2010.pdf [07.10.2016]
- Rubie-Davies, C. M., Flint, A., & McDonald, L. G. (2011).** Teacher beliefs, teacher characteristics, and school contextual factors: what are the relationships? *British Journal of Educational Psychology*, 82(2), 270-288.
- Ruf, U. & Gallin, P. (1991).** Lernen auf eigenen Wegen - mit Kernideen und Reisetagebüchern [Learning on own paths – with core ideas and itineraries]. *Beiträge zur Lehrerbildung*, 9(2), 248-258.
- Ruiz-Primo, M. A. & Furtak, E. M. (2007).** Exploring teacher's informal formative assessment practices and student's understanding of the context of scientific inquiry. *Journal of Research in Science Teaching*, 44, 57-84.
- Ruiz-Primo, M. A., Furtak, E. M., Ayala, C., Yin, Y., & Shavelson, R.J. (2010).** Formative assessment, motivation, and science learning. In H. Andrade & G.J. Cizek (Ed.), *Handbook of formative assessment* (pp. 139 – 158). New York: Routhledge.
- Ruiz-Primo, M. A. & Li, M. (2013).** Analysing teachers' feedback practices in response to students' work in science classrooms. *Applied Measurement in Education*, 26(3), 163 – 175.
- Sadler, D. R. (1989).** Formative assessment and the design of instructional systems. *Instructional Science*, 18(2), 119–144.
- Santos, L., & Dias, S. (2006).** Como entendem os alunos o que lhes dizem os professores? A complexidade do feedback [How do the students understand what the teachers say? On the complexity of feedback]. *ProfMat 2006*. Lisboa: APM.
- Sato, M., Wei, R.C., & Darling-Hammond, L. (2008).** Improving teachers' assessment practices through professional development: The case of national board certification. *American Educational Research Journal*, 45(3), 669 – 700.
- Schecker, H., Fischer, H. E., & Wiesner, H. (2004).** Physikunterricht in der gymnasialen Oberstufe [Physics education at upper secondary school]. In H.-E. Tenorth (Ed.), *Kerncurriculum Oberstufe II. Biologie, Chemie, Physik, Geschichte, Politik* (pp. 148–234). Weinheim: Beltz.
- Schmitt, N. (1996).** Uses and abuses of coefficient alpha. *Psychological Assessment*, 8(4), 350–353.
- Schunk, D. (2003).** Self-efficacy for reading and writing: Influence of modelling, goal-setting, and self-evaluation. *Reading & Writing Quarterly*, 19, 159–172.
- Schwartz, M., (1984).** Response to writing: A college-wide perspective. *College English*, 46, 55–62.

- Schwartz, G. & Allal, L. (2000).** Vers une pratique de l'évaluation formative dans le secondaire I. Analyses d'expériences menées au cycle d'orientation de Genève [Towards formative assessment practices at lower secondary school level. Analyses of trials at the orientation term in Geneva]. DIPCO: Genève.
- Schwarz, C. V. & White, B. Y. (2005).** Metamodeling knowledge: Developing students' understanding of scientific modelling. *Cognition and Instruction*, 23(2), 165–205.
- Scriven, M. (1967).** The methodology of evaluation. In R. W. Tyler, R. M. Gagné, & M. Scriven (Eds.), *Monograph Series on Educational Evaluation: Vol. 1. Perspectives of curriculum evaluation* (pp. 39–83). Chicago: Rand McNally.
- Shavelson, R. J., Young, D. B., Ayala, C. C., Brandon, P. R., Furtak, E. M., Ruiz-Primo, M. A., Tomita, M. K., & Yin, Y. (2008).** On the impact of curriculum-embedded formative assessment on learning: A collaboration between curriculum and assessment developers. *Applied Measurement in Education*, 21(4), 295–314.
- Shell, D. F., Colvin, C., & Brunning, R. H. (1995).** Self-efficacy, attributions, and outcome expectancy mechanisms in reading and writing achievement: Grade-level and achievement level differences. *Journal of Educational Psychology*, 87, 386 – 398.
- Shepard, L. (2000).** The role of assessment in a learning culture. *Educational Researcher*, 29(7), 4–14.
- Shute, V.J. (2008).** Focus on formative feedback. *Review of educational research*, 78(1), 153 – 189.
- Singer, J., Marx, R. W., Krajcik, J. S., & Chambers, J. C. (2000).** Constructing extended inquiry projects: Curriculum materials for science education reform. *Educational Psychologist*, 35(3), 165–178.
- Sluismans, D. M. A. (2002).** Student involvement in assessment, the training of peer-assessment skills. Maastricht: Interuniversity Centre for Educational Research.
Retrieved from <https://www.ou.nl> [07.10.2016]
- Smit, R. (2009).** Die formative Beurteilung und ihr Nutzen für die Entwicklung von Lernkompetenz. Eine empirische Studie in der Sekundarstufe 1 [Formative assessment and its use for the development of learning competence. An empirical study at lower secondary school level]. Schneider Verlag Hohengehren GmbH: Baltmannsweiler.
- Smit, R. & Birri, T. (2012).** Lernen mit Rubrics als Teil der formativen, standardorientierten Beurteilung [Learning with rubrics as part of formative, standard-oriented assessment]. Unpublished manuscript, PH St. Gallen.
- Smit, R. & Birri, T. (2014).** Assuring the quality of standards-oriented classroom assessment with rubrics for complex competencies. *Studies in Educational Evaluation*, 43, 5-13.
- Smith, E. & Gorard, S. (2005).** 'They don't give us our marks': The role of formative feedback in student progress. *Assessment in Education: Principles, Policy & Practice*, 12, 21–38.
- So, W.W.M. & Lee, T.T.H. (2011).** Influence of teachers' perceptions of teaching and learning on the implementation of assessment for learning in inquiry study. *Assessment in Education: Principles, Policy and Practice*, 18(4), 417 – 432.
- Stallings, V. & Tascione, C. (1996).** Student self-assessment and self-evaluation. *Mathematics Teacher*, 89, 548–55.
- Stern, L. A. & Solomon, A., (2006).** Effective faculty feedback: The road less travelled. *Assessing Writing*, 11, 22–41.
- Stiggins, R. J. (1999).** Evaluating classroom assessment training in teacher education programs. *Educational Measurement: Issues and Practice*, 18(1), 23-27.
- Stiggins, R. J. (2005).** Student-involved assessment for learning. Upper Saddle River, NJ: Prentice Hall.
- Stiggins, R. J., Griswold, M. M. & Wiklund, K. R. (1989).** Measuring thinking skills through classroom assessment. *Journal of Educational Measurement*, 26, 233-246.
- Stracke E. & Kumar, V. (2010).** Feedback and self-regulated learning: Insights from supervisors' and Ph.D. examiners' reports. *Reflective Practice*, 11(1), 19–32.
- Swaffield, S., (Ed.), (1998).** Unlocking assessment. Understanding for reflection and application. London / New York: Routledge.
- Thompson, A. G. (1992).** Teachers' beliefs and conceptions: A synthesis of the research. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 127-146). New York: Macmillan.
- Tierney, R. D. (2006).** Changing practices: Influences on classroom assessment. *Assessment in Education: Principles, Policy & Practice*, 13, 239–264.

- Topping, K. (1998).** Peer assessment between students in colleges and universities. *Review of Educational Research*, 68, 249–276.
- Topping, K. J. (2003).** Self and peer assessment in school and university: Reliability, validity and utility. In M. Segers, F. Dochy & E. Cascaller (Eds.), *Optimising new modes of assessment: in search of qualities and standards* (pp. 55–87). The Netherlands: Kluwer Academic Publishers.
- Topping, K. J. (2010).** Peers as a source of formative assessment. In H. Andrade & G.J. Cizek (Eds.), *Handbook of formative assessment* (pp. 61-74). New York: Routledge.
- Topping, K., Smith, F. F., Swanson, I., & Elliot, A. (2000).** Formative peer assessment of academic writing between postgraduate students. *Assessment and Evaluation in Higher Education*, 25, 149-169.
- Tsivitanidou, O. & Labudde, P. (2016).** Can peer-assessment inform teachers about secondary-school students' mastery of the modelling competence? In J. Lavonen, K. Juuti, J. Lampiselkä, A. Uitto, & K. Hahl (Eds.), *Science Education Research: Engaging learners for a sustainable future. Proceedings of ESERA 2015* (pp. 1619 – 1630). Helsinki: University of Helsinki.
- Tsivitanidou, O., Zacharia, Z. C. & Hovardas, A. (2011).** High school students' unmediated potential to assess peers: Unstructured and reciprocal peer assessment of web-portfolios in a science course. *Learning and Instruction*, 21, 506-519.
- Tuck, J. (2012).** Feedback-giving as social practice: Teachers' perspectives on feedback as institutional requirement, work and dialogue. *Teaching in higher education*, 17(2), 209 – 221.
- Vögeli - Mantovani, U. (1999).** SKBF Trendbericht Nr. 3: Mehr fördern, weniger auslesen. Zur Entwicklung der schulischen Beurteilung in der Schweiz [SKBF trend report no. 3: More support, less selection. On the development of educational assessment in Switzerland]. Aarau: Schweizerische Koordinationsstelle für Bildungsforschung.
- Walker, M. (2009).** An investigation into written comments on assignments: Do students find them usable? *Assessment & Evaluation in Higher Education*, 34(1), 67-78.
- Watson, A. (2006).** Some difficulties in informal assessment in mathematics. *Assessment in Education: Principles, Policy & Practice*, 13(3), 289–303.
- Weaver, M. (2006).** Do students value feedback? Student perceptions of tutors' written responses. *Assessment and Evaluation in Higher Education*, 31, 379–394.
- Weinert, F. E. (2001).** Vergleichende Leistungsmessung in Schulen – eine umstrittene Selbstverständlichkeit [Comparative measurement of student achievement at schools – a controversial obviousness]. In F.E. Weinert (Ed.), *Leistungsmessungen in Schulen* (pp. 17-31). Weinheim, Basel, Bonn: Beltz.
- Welch, W. W., Klopfer, L.E., Aikenhead, G. S., & Robinson, J.T. (1981).** The role of inquiry in science education: Analysis and recommendations. *Science Education*, 65(1), 33-50.
- White, B. Y. & Frederiksen, J. R. (1998).** Inquiry, modelling, and metacognition: Making science accessible to all students. *Cognition and Instruction*, 16(1), 3–118.
- Widmer Märki, I. (2011).** Fächerübergreifender naturwissenschaftlicher Unterricht: Umsetzung und Beurteilung von Schülerleistungen im Gymnasium [Interdisciplinary science education: Implementation and assessment of student achievement at the gymnasium]. PhD Thesis. Basel: University of Basel.
- Wiggins, G.P. (1998).** *Educative assessment: Designing assessments to inform and improve student performance*. San Francisco, CA: Jossey-Bass.
- Wilcox, B. L. (1997).** Writing portfolios: Active vs. passive. *English Journal*, 86, 7-34.
- Wilcoxon, F. (1945).** Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6), 80–83.
- Wiliam, D. (2006).** Formative assessment: Getting the focus right. *Educational Assessment*, 11(3-4), 283–289.
- Wiliam, D. (2010).** Research literature and implications for a new theory of formative assessment. In H. Andrade & G.J. Cizek (Eds.), *Handbook of formative assessment* (pp. 18-40). New York: Routledge.
- Wiliam, D., Lee, C., Harrison, C., & Black, P. (2004).** Teachers developing assessment for learning: Impact on student achievement. *Assessment in Education: Principles, Policy & Practice*, 11, 49–65.
- Williams, J. & Ryan, J. (2000).** National testing and the improvement of classroom teaching: Can they co-exist? *British Educational Research Journal*, 26(1), 49–73.
- Wilson, M., & Sloane, K. (2000).** From principles to practice: An embedded assessment system. *Applied Measurement in Education*, 13(2), 181–208.

- Windschitl, M. (2004).** Folk theories of 'inquiry': How preservice teachers reproduce the discourse and practices of a theoretical scientific method. *Journal of Research in Science Teaching*, 41(5), 481–512.
- Woolfolk Hoy, A., Davis, H., & Pape, S. J. (2006).** Teacher knowledge and beliefs. In P. A. Alexander, & P. H. Winne (Eds.), *Handbook of educational psychology* (pp. 715 - 737). Mahwah, NJ: LEA.
- Yancey, K. B. (1998).** Getting beyond exhaustion: Reflection, self-assessment, and learning. *Clearing House*, 72, 13–17.
- Yin Y., Shavelson, R.J., Ayala, C.C., Araceli Ruiz-Primo, M., Brandon, P. R., Furtak, E. M., Tomita, M. K. & Young, D. B. (2008).** On the impact of formative assessment on student motivation, achievement, and conceptual change. *Applied Measurement in Education*, 21(4), 335-359.
- Yoon, S., Pedretti, E., Pedretti, L., Hewitt, J., Perris, K., & Van Oostveen, R. (2006).** Exploring the use of cases and case methods in influencing elementary teachers' self-efficacy beliefs. *Journal of Science Teacher Education*, 17, 15–35.
- Zachos, P., Hick, T. L, Doane, W. E. J., & Sargent, C. (2000).** Setting theoretical and empirical foundations for assessing scientific inquiry and discovery in educational programs. *Journal of Research in Science Teaching*, 37(9), 938 – 962.
- Zaborowski, K. U., Meier, M., & Breidenstein, G. (2011).** Leistungsbewertung und Unterricht - Ethnographische Studien zur Bewertungspraxis in Gymnasium und Sekundarschule [Assessment of student achievement and education – ethnographic studies on the practice of assessment at lower and upper secondary school level]. Wiesbaden: VS- Verlag.
- Zimmerman, B. & Schunk, D. (2004).** Self-regulating intellectual processes and outcomes: A social cognitive perspective. In D. Dai & R. Sternberg (Eds.), *Motivation, emotion, and cognition: Integrative perspectives on intellectual functioning and development* (pp. 323–349). Mahwah, NJ: Lawrence Erlbaum Associates.

Appendix

A1. Teacher profile questionnaire

1: [Introduction to questionnaire]	
2: What is your name?	
3: How many years of teaching experience do you have?	3a: 0-4 3b: 5-10 3c: 11-20 3d: More than 20
4: Which subject or subjects do you teach?	4a: Physics 4b: Chemistry 4c: Biology 4d: Technology 4e: Mathematics 4f: Integrated science 4g: Other: 4h: Other:
5: Which school level or levels do you teach at?	5a: 4.-6. grade (years 10-12) 5b: 7.-9. grade (years 13-15) 5c: 10.-12. grade (years 16-18) 5d: Other:
6: When you think of your educational background, how much of it has been pedagogical?	6a: 0-25% 6b: 26-50% 6c: 51-75% 6d: 76-100%

	15: How often do you...	16: How important do you think it is to...	17: How competent are you to...
18: ...assess student learning to decide on the student's next learning step?	15a: In all lessons 15b: Once a week 15c: Once every second week 15d: Once per month 15e: Less than once per month 15f: I don't know	16a: Very important 16b: Not important 16c: I don't know	17a: Very competent 17b: Not competent 17c: I don't know
19: ...refer to assessment criteria in your formative feedback?			
20: ...communicate students' progress to them through oral feedback?			
21: ...communicate students' progress to them through written feedback?			
22: ...communicate students' progress to them through marked tests or marked assignments?			
23: ...communicate students' progress to their parents?			
24: ...adapt your own teaching based on the results of the formative feedback?			

	25: How often do...	26: How important do you think it is to let...	27: How competent are you in having...
28: ...students in your class assess their own work?	25a: In all lessons 25b: Once a week	26a: Very important 26b: Not important	27a: Very competent 27b: Not competent
29: ...students in your class assess the work of their peers?	25c: Once every second week 25d: Once per month 25e: Less than once per month 25f: I don't know	26c: I don't know	27c: I don't know

	30: How often do you...	31: How important do you think it is to...	32: How competent are you to...
33: ...assess the students in your class summatively?	30a: Once a week or more	31a: Very important 31b: Not important	32a: Very competent 32b: Not competent
34: ...use tests or quizzes as summative assessment?	30b: 1-3 times per month 30c: Every second month	31c: I don't know	32c: I don't know
35: ...involve the students in your class in the summative assessment?	30d: 1-2 times per semester 30e: Less than once per semester 30f: I don't know		

36: Are you the one who designs the summative assessment tools/procedures for the students in your class?	36a: Yes 36b: No 36c: Some of them 36d: I don't know
--	---

<p>37: Please indicate the degree to which you agree or disagree with each statement below. For your answers, think of when you will use formative assessment within a teaching unit based on inquiry. (Modified from Enochs, L., & Riggs, I. (1990). <i>School Science and Mathematics</i>, 90, 694-706.)</p>	<p>37a: Strongly agree 37b: Strongly disagree 37c: I don't know</p>
<p>38: I will continually find better ways to teach using formative assessment</p>	
<p>39: Even if I try very hard, it will be difficult for me to integrate formative assessment into my teaching</p>	
<p>40: I know the steps necessary to teach effectively using formative assessment</p>	
<p>41: I will not be very effective in monitoring student work when I teach using formative assessment</p>	
<p>42: My teaching will not be very effective when using formative assessment</p>	
<p>43: The inadequacy of a student's background can be overcome by the use of formative assessment</p>	
<p>44: When a low-achieving student progresses, it is usually due to formative assessment given by the teacher</p>	
<p>45: I understand formative assessment well enough to be effective using it</p>	
<p>46: Increased effort of the teacher in using formative assessment produces little change in some students' achievement in inquiry based competences</p>	
<p>47: When using formative assessment, I will find it difficult to explain subject content to students</p>	
<p>48: I will typically be able to answer students' questions when using formative assessment</p>	
<p>49: I wonder if I will have the necessary skills to use formative assessment</p>	

A3. Evaluation form for teachers

Evaluationsformular zur erprobten formativen Beurteilungsmethode

Name: _____

Bitte lege Deine Unterrichtsvorbereitung (Präp), allfällige Arbeitsblätter, Arbeitsaufträge etc. bei.

1) In welcher Klasse hast Du die Unterrichtseinheit (UE) mit formativer Beurteilung eingesetzt?

Klassenstufe (Primarschule: alte Zählweise, Klasse 1-9)			
Fach			
Anzahl Schülerinnen		Anzahl Schüler	

2) In welcher konkreten UE hast Du formativ beurteilt? Über wie viele Lektionen erstreckt sich Deine Einheit?

3) Auf welche formative Beurteilungsmethode beziehen sich Deine Antworten? Wenn Du mehr als eine Beurteilungsmethode erprobt hast, gib hier bitte trotzdem nur **eine** an und beziehe Dich bei den weiteren Fragen ausschliesslich darauf.

Beurteilungsmethode Bitte setze ein Kreuz pro Spalte.	Konkrete Umsetzung Bitte setze ein Kreuz pro Spalte.
<input type="checkbox"/> Schriftliche Rückmeldungen durch die LP	<input type="checkbox"/> Mit vorgedruckten Kriterienrastern <input type="checkbox"/> Mit offenen schriftlichen Kommentaren <input type="checkbox"/> Anderes: _____
<input type="checkbox"/> Selbst- und Partnerbeurteilungen	<input type="checkbox"/> Selbstbeurteilung (entspricht Reflexion) <input type="checkbox"/> Partnerbeurteilung (SuS beurteilen sich paarweise gegenseitig) <input type="checkbox"/> Klassenbeurteilung (alle SuS beurteilen alle ausgelegten Arbeiten) <input type="checkbox"/> Anderes: _____
<input type="checkbox"/> Offene und strukturierte Diskussionen in der ganzen Klasse	<input type="checkbox"/> Offene Diskussion (Verlauf der Diskussion durch die Inhalte bestimmt) <input type="checkbox"/> Strukturierte Diskussion (Verlauf der Diskussion formal in Phasen unterteilt; SuS haben Rollen) <input type="checkbox"/> Anderes: _____
<input type="checkbox"/> Andere	

4) Auf welche Kompetenzen und Teilkompetenzen bezieht sich die gewählte formative Beurteilung?

Kompetenz	Teilkompetenzen (siehe Seite 15 - 18 im Manual)
<input type="checkbox"/> Experimentieren und Untersuchen	
<input type="checkbox"/> Konstruieren	
<input type="checkbox"/> Argumentieren	
<input type="checkbox"/> Mit Modellen arbeiten	
<input type="checkbox"/> Innovativ denken und handeln	
<input type="checkbox"/> andere	

5) Wie planst Du den Unterrichtsablauf der UE? Bitte gehe ausdrücklich auf die Sequenz mit formativer Beurteilung ein.

Statt die untenstehende Tabelle auszufüllen kannst Du gerne auch Deine ganz normale "Präp" kopieren und beilegen.

Dauer (Min)	Unterrichtsinhalt	Aktivitäten der LP und Aktivitäten der SuS

6) Allgemein betrachtet: Welche Vorteile siehst Du bei der in Frage 3 angegebenen Beurteilungsmethode?

7) Allgemein betrachtet: Welche Nachteile siehst Du bei der in Frage 3 angegebenen Beurteilungsmethode?

8) In der konkreten Situation: Welche Vorteile hast Du beim Erproben der Beurteilungsmethode erkannt?

[Empty response box for question 8]

9) In der konkreten Situation: Welche Schwierigkeiten hast Du beim Erproben der Beurteilungsmethode erkannt?

[Empty response box for question 9]

10) Wie, mit welchen Hilfestellungen, könnten LPs diese Schwierigkeiten überwinden?

[Empty response box for question 10]

11) Hat sich die formative Beurteilung für die SuS gelohnt? In welcher Hinsicht (nicht)?

[Empty response box for question 11]

12) Hat sich die formative Beurteilung für Dich als Lehrperson gelohnt? In welcher Hinsicht (nicht)?

[Empty response box for question 12]

13) Was würdest Du in Bezug auf Ablauf und Gestaltung verändern, wenn Du diese Beurteilungsmethode in der gleichen Unterrichtseinheit erneut einsetzen würdest?

[Empty response box for question 13]

14) Bitte kreuze an, in wie weit Du den Aussagen in der Tabelle zustimmst. Beziehe Dich bitte auf die gleiche Beurteilungsmethode wie in Frage 3.

	1 Stimme überhaupt nicht zu	2 Stimme eher nicht zu	3 Stimme eher zu	4 Stimme völlig zu
Die Beurteilungsmethode war im Unterricht einfach umsetzbar.				
Die Beurteilungsmethode erlaubte auf einfache Weise, die Schwierigkeiten und Bedürfnisse der SuS zu identifizieren.				
Die Beurteilungsmethode erlaubte auf einfache Weise, hilfreiche Kommentare für die SuS zu formulieren.				
Die Beurteilungsmethode erforderte mehr Zeit, als ich erwartet hatte.				
Die Beurteilungsmethode war für die SuS nützlich.				
Die Beurteilungsmethode war für mich als LP nützlich.				
Die Beurteilungsmethode ist Teil meines Unterrichtsverständnisses.				
<i>Falls "stimme überhaupt nicht zu" oder "stimme eher nicht zu" angekreuzt:</i> Die Beurteilungsmethode könnte ich leicht in mein Unterrichtsverständnis integrieren.				
Die Beurteilungsmethode nutze ich in meiner normalen Unterrichtspraxis.				
<i>Falls "stimme überhaupt nicht zu" oder "stimme eher nicht zu" angekreuzt:</i> Die Beurteilungsmethode könnte ich leicht in meiner normalen Unterrichtspraxis nutzen.				
Die Beurteilungsmethode ist Bestandteil der normalen Unterrichtspraxis in der Schweiz.				
Die Beurteilungsmethode ist auch für summative Zwecke einsetzbar.				
Die Beurteilungsmethode kann mir hilfreiche Informationen für meine nächsten Unterrichtsschritte liefern.				
Die SuS sind gewohnt, mit dieser Beurteilungsmethode zu arbeiten.				
Die SuS waren in der Lage, die Rückmeldungen, welche sie erhalten hatten, zu interpretieren.				
<i>Nur bei Partner- und Selbstbeurteilung:</i> Die SuS waren motiviert, einander Rückmeldungen zu geben / zu reflektieren.				
Die SuS waren motiviert, sich mit den erhaltenen Rückmeldungen auseinanderzusetzen.				
Die SuS waren motiviert, die Rückmeldungen in ihre weitere Arbeit einzubeziehen.				

Ganz herzlichen Dank für Dein "Mitdenken" und Deine Unterstützung im Projekt!

A4. Evaluation form for students

Liebe Schülerin, lieber Schüler,




in diesem Semester haben Sie im Physikpraktikum Kommentare in die Protokolle von MitschülerInnen geschrieben und auch Kommentare zu Ihren eigenen Protokollen erhalten: im Praktikum „Schaltkreise“

Bitte beantworten Sie zu diesen Praktika die folgenden Fragen. Geben Sie – ausser bei Frage 1 – jeweils eine Antwort für den Aspekt „Kommentare schreiben“ (linke Spalte) und eine Antwort für den Aspekt „Kommentare erhalten“ (rechte Spalte) ab. Ihre Antworten werden vertraulich behandelt.

1) Wie <u>unterscheiden</u> sich die Kommentare Ihrer MitschülerInnen von den Kommentaren, welche Herr Scandella normalerweise in die Protokolle hineinschreibt?	
2) Bitte kreuzen Sie eine Antwort an. Wie <u>hilfreich</u> ...	
<p>... war es für Sie, die Protokolle von MitschülerInnen zu kommentieren?</p> <p><input type="checkbox"/> sehr hilfreich</p> <p><input type="checkbox"/> eher hilfreich</p> <p><input type="checkbox"/> eher nicht hilfreich</p> <p><input type="checkbox"/> gar nicht hilfreich</p>	<p>... waren die Kommentare, die Sie von MitschülerInnen erhalten haben, für Sie?</p> <p><input type="checkbox"/> sehr hilfreich</p> <p><input type="checkbox"/> eher hilfreich</p> <p><input type="checkbox"/> eher nicht hilfreich</p> <p><input type="checkbox"/> gar nicht hilfreich</p>
3) Bitte begründen Sie Ihre Antwort auf Frage 2.	
4) <u>Was</u> haben Sie ...	
<p>... beim Kommentieren der Protokolle Ihrer MitschülerInnen <u>gelernt</u>? Allenfalls können Sie auch die Erfahrungen aufschreiben, die Sie gemacht haben.</p>	<p>... aus den Kommentaren der MitschülerInnen <u>gelernt</u>?</p>

A5. Interview questions for teachers

	Fachhochschule Nordwestschweiz Pädagogische Hochschule		
Leitfaden Einzelinterview mit Lehrpersonen			
Einleitung Ich möchte Dir herzlich danken, dass Du Dich bereit erklärt hast, an diesem Gespräch teilzunehmen. Es ist für mich sehr nützlich, wenn ich in dieser frühen Phase der Zusammenarbeit einen möglichst guten Einblick erhalte, wie Du und die anderen am Projekt beteiligten Lehrpersonen arbeiten.			
Grundlage dieses Gesprächs ist Deine Erprobung einer formativen Beurteilungsmethode im Rahmen des Projekts ASSIST-ME. Du hast dazu bereits ein Evaluationsformular ausgefüllt, jetzt geht es darum, Deine Erfahrungen, Einschätzungen und Rückmeldungen zu der Beurteilungsmethode aber auch zu formativer Beurteilung ganz allgemein zu vertiefen.			
Das Gespräch wird etwa 30, maximal 40 Minuten dauern. Du sprichst hier nur für Dich selbst. Was Du erzählst, wird vertraulich behandelt. Auch wenn ich natürlich Deinen Namen kenne, werden in den Berichten und Texten keine Namen erscheinen.			
Das Interview folgt einem standardisierten Verfahren. Das bedeutet konkret, dass ich oft nichts zu Deinen Antworten sage, weil ich Deine Meinung nicht beeinflussen will. Im Gespräch wirkt das möglicherweise unnatürlich.			
Ich werde das Gespräch aufnehmen, um es anschliessend auswerten zu können. Falls Du Fragen hast oder ich mich unklar ausdrücke, bitte ich Dich, einfach zurückzufragen. Hast Du jetzt gerade Fragen? Dann beginnen wir.			
Institut Forschung & Entwicklung Zentrum Naturwissenschafts- und Technikdidaktik	Riehenstrasse 154 CH-4058 Basel	www.fhnw.ch www.assistme.ku.dk	Seite 1

	Fachhochschule Nordwestschweiz Pädagogische Hochschule		
A Allgemeine Angaben zur Unterrichtseinheit, in der die formative Beurteilungsmethode erprobt wurde			
1	In welchem Unterrichtsthema hast Du formativ beurteilt?		
2	Mit welcher formativen Beurteilungsmethode hast Du gearbeitet?		
3	Würdest Du bitte die Unterrichtseinheit zusammenfassen und dabei auch auf die formative Beurteilung eingehen? / In welcher Phase der Unterrichtseinheit hast Du die formative Beurteilung eingebaut? / Warum hast Du die formative Beurteilung gerade da eingebaut?		
4	Welchen Kompetenzbereich hast du in die Beurteilung mit einbezogen? / Welche Teilkompetenzen hast du in die Beurteilung mit einbezogen?		
B Zur Vorbereitung			
5	Als es darum ging, eine formative Beurteilungsmethode im Unterricht zu erproben, wie hast Du mit der Planung angefangen? (Elemente: Inhalt, Klasse, formative Beurteilungsmethode, Kompetenzbereich)		
6	Warum hast Du gerade diese Beurteilungsmethode zur Erprobung ausgewählt?		
7	Warum hast Du die formative Beurteilung gerade in diesem Themenbereich ausprobiert?		
8	Aus welchem Grund hast Du Dich für diesen Kompetenzbereich entschieden?		
9	Warum hast Du die formative Beurteilung gerade in dieser Klasse ausprobiert?		
10	Hattest Du bei der Vorbereitung bestimmte Befürchtungen oder Erwartungen hinsichtlich des Einsatzes der formativen Beurteilungsmethode?		
C Zur Durchführung			
11	Was lief bei der Umsetzung der Unterrichtseinheit so wie geplant?		
12	Was lief bei der Umsetzung der Unterrichtseinheit anders als geplant? (beispielsweise in Bezug auf Zeitplanung, Engagement der SuS, Verständnis der Beurteilungskriterien, inhaltliche Änderungen, ...)		
13	Auf welche Handlungen oder Produkte der Schülerinnen und Schüler (wie beispielsweise mündliche Äusserungen, Einträge im Laborjournal, ...) war die formative Beurteilung abgestützt? / Warum? / Waren diese Handlungen und Produkte geeignet, die SuS formative beurteilen?		
14	Du hast gesagt, Deine Beurteilung habe sich auf den Kompetenzbereich ... (vgl. Frage 4) bezogen. Wie hat sich das konkret in der Unterrichtseinheit geäussert?		
15	Nur bei Selbst- und Partnerbeurteilung: Waren die SuS darin geübt? Haben sie sich auch schon gegenseitig Rückmeldungen gegeben?		
16	Gab es explizite Beurteilungskriterien? Warum? Wie wurden die Beurteilungskriterien entwickelt, kommuniziert oder diskutiert?		
Institut Forschung & Entwicklung Zentrum Naturwissenschafts- und Technikdidaktik	Riehenstrasse 154 CH-4058 Basel	www.fhnw.ch www.assistme.ku.dk	Seite 2

D	Evaluation der erprobten Beurteilungsmethode und der Methodenbeschreibung im Manual
17	Unabhängig von der Unterrichtseinheit: würdest Du die Beurteilungsmethode nochmals einsetzen? / Warum (nicht)?
18	Was würdest Du verändern, wenn Du nächstes Semester die gleiche Beurteilungsmethode noch einmal einsetzen würdest? (in Bezug auf: Produkte und Handlungen der SuS, Einbezug der SuS, Beurteilungskriterien, Aufgabenstellung. Bsp: Zeitaufwand)
19	Wo liegen die Stärken und die Schwächen der Beurteilungsmethode, die Du ausprobiert hast?
20	In wie weit ist die Beurteilungsmethode für andere Schulstufen und für andere Fächer geeignet?
E	Generelle Einschätzungen zu formativer Beurteilung
21	Welche generellen Chancen und Probleme siehst Du im Zusammenhang mit formativer Beurteilung auf Deiner Schulstufe?
22	Was ist wichtig bei der formativen Beurteilung auf Deiner Schulstufe, worauf muss man achten?
23	Im Projekt wurden spezifische Kompetenzbereiche beschrieben, welche mit forschend-entdeckendem Lernen im Zusammenhang stehen. Welche der Kompetenzen liessen sich mit welchen der drei Beurteilungsmethoden (vgl. Raster) kombinieren?
24	Welche Art von Unterstützung würdest Du Dir wünschen, um FA in Deinen Unterricht einzuflechten? <i>Spezifisch zu formativer Beurteilung beim forschend-entdeckenden Lernen: zu welchen Zeitpunkten/bei welchen Sequenzen innerhalb von forschend-entdeckenden Unterrichtseinheiten ist formative Beurteilung besonders nützlich?</i>
25	Bei der Umsetzung der formativen Beurteilungsmethoden in diesem Projekt: Welche Aspekte der Arbeit im ASSIST-ME Projekt haben Dir am meisten geholfen (bspw. Methodenbeschreibungen, Kompetenzbeschreibungen, Beurteilungskriterien und Rubrics, paradigmatische Beispiele, Diskussion mit anderen Lehrpersonen, ...)? Gründe?
26	Welche Verbesserungsvorschläge hast Du fürs Manual?
27	Das war die letzte Frage. Möchtest Du noch etwas sagen oder los werden, was bisher nicht zur Sprache gekommen ist?

So, das ist das Ende des Interviews. Hast Du jetzt noch Fragen?

A6. Topics for teacher group discussions

Group discussion after 1st round of implementation (7th January 2015)

- 1) How did you put „your“ assessment method into practice?
- 2) What benefits and what challenges did you perceive?
- 3) What hints, tips and tricks, problems, ... relating to the formative assessment method used should be mentioned regarding the work in the next semester?
- 4) What general questions and problems related to formative assessment have emerged during the work in the last months?

Group discussion after 2nd round of implementation (27th May 2015)

- 1) How did you put „your“ assessment method into practice?
- 2) What benefits and what challenges did you perceive?
- 3) What measures have supported the successful use of the formative assessment method (in terms of contents, methods, ...)?

Group discussions after 3rd round of implementation (5th January 2016)

- 1) How did you put „your“ assessment method into practice?
- 2) What benefits and what challenges did you perceive?
- 3) What are the main challenges in terms of formative assessment?

A7. Description of cases

Round	School level	Teacher	Criteria related to inquiry units (see section 5.4.3)			Criteria related to formal formative assessment (see section 5.4.4)				Documentation of cases (see section 5.4.2)
			Inquiry activity	Dimension(s) of openness	assessed competence(s)	Communication of criteria	Data for diagnosis	Formative assessment method(s)	Use of feedback	
1	Primary 3 rd grade	P1	Exploring buoyancy	<ul style="list-style-type: none"> • Methods • Solution processes • Solutions 	<ul style="list-style-type: none"> • Hypothesis generation • Investigation • Analysis and interpretation • Communication 	Aims were introduced to the students in a written form at the beginning of the unit	<ul style="list-style-type: none"> • Observation of students' performance • Presentation 	<ul style="list-style-type: none"> • Peer-assessment (written, structured by questions) 	<ul style="list-style-type: none"> • Transfer to very similar activity in the second part of the unit 	<ul style="list-style-type: none"> • Teacher evaluation form • Teaching material • Group discussion
	Primary 5 th grade	P2	Exploring buoyancy	<ul style="list-style-type: none"> • Methods • Solution processes • Solutions 	<ul style="list-style-type: none"> • Hypothesis generation • Investigation • Analysis and interpretation • Communication 	Aims were introduced to the students in a written form at the beginning of the unit	<ul style="list-style-type: none"> • Observation of students' performance • Presentation 	<ul style="list-style-type: none"> • Peer-assessment (written, structured by questions) 	<ul style="list-style-type: none"> • Transfer to very similar activity in the second part of the unit 	<ul style="list-style-type: none"> • Teacher evaluation form • Teaching material • Group discussion
	Primary 4 th grade	P3	Constructing a model to explain astronomical phenomena	<ul style="list-style-type: none"> • Strategy • Methods • solutions • Solution processes 	<ul style="list-style-type: none"> • Investigation • Interact in heterogeneous groups 	Criteria were introduced to the students in a written form at the beginning of the unit	<ul style="list-style-type: none"> • Observation of students' performance • Model 	<ul style="list-style-type: none"> • Peer-assessment (written; rubrics and open comments; focussing on interaction) • Written teacher assessment (rubrics and open comments; focussing on investigation) 	<ul style="list-style-type: none"> • Revision of draft artefact for domain-specific competence • Transfer to later situations for transversal competences 	<ul style="list-style-type: none"> • Teacher evaluation form • Teaching material
	Primary 4 th grade	P4	Exploring and presenting information on human organs	<ul style="list-style-type: none"> • Content • Methods • Solutions • Solution processes 	<ul style="list-style-type: none"> • Communication 	Aims were introduced to the students in a written form at the beginning of the unit	<ul style="list-style-type: none"> • Presentation 	<ul style="list-style-type: none"> • Peer-assessment (written; rubrics) 	<ul style="list-style-type: none"> • None / unclear 	<ul style="list-style-type: none"> • Teacher evaluation form • Teaching material • Group discussion • Individual interview
	Primary 3 rd grade	P5	Constructing a car	<ul style="list-style-type: none"> • Strategy • Solutions • Solution processes 	<ul style="list-style-type: none"> • Investigation 	Aims were introduced to the students in a written form at the beginning of the unit	<ul style="list-style-type: none"> • Constructed car 	<ul style="list-style-type: none"> • Written teacher assessment (open comments) 	<ul style="list-style-type: none"> • Revision of draft version 	<ul style="list-style-type: none"> • Teacher evaluation form • Teaching material • Group discussion
	Primary 4 th grade	P6	Constructing a pendulum clock	<ul style="list-style-type: none"> • Solution processes 	<ul style="list-style-type: none"> • Planning • Investigation 	Criteria were introduced to the students in a written form at the beginning of the unit	<ul style="list-style-type: none"> • Reflections on construction process 	<ul style="list-style-type: none"> • Self-assessment (oral, in groups) 	<ul style="list-style-type: none"> • Revision of draft version 	<ul style="list-style-type: none"> • Teacher evaluation form • Teaching material • Group discussion • Individual interview
	Primary 4 th grade	P7	Mixing different substances	<ul style="list-style-type: none"> • Methods • Solution processes 	<ul style="list-style-type: none"> • Investigation • Interact in heterogeneous groups • Act autonomously 	Aims were introduced to the students in a written form at the beginning of the unit	<ul style="list-style-type: none"> • Observation of students' performance 	<ul style="list-style-type: none"> • Written teacher assessment (rubrics and open comments) 	<ul style="list-style-type: none"> • Transfer to later activities and situations 	<ul style="list-style-type: none"> • Teacher evaluation form • Teaching material
	Primary 3 rd grade	P8	Investigating how animals move	<ul style="list-style-type: none"> • Solutions • Solution processes 	<ul style="list-style-type: none"> • Investigation • Communication • Interact in heterogeneous groups • Act autonomously 	Aims were introduced to the students in a written form at the beginning of the unit	<ul style="list-style-type: none"> • Observation of students' performance • Written report 	<ul style="list-style-type: none"> • Peer-assessment (oral, focussing on transversal competences) • Written teacher assessment (focussing on domain-specific competences) 	<ul style="list-style-type: none"> • Revision of draft report for domain-specific competences • Transfer to later situations for transversal competences 	<ul style="list-style-type: none"> • Teacher evaluation form • Teaching material
	Primary 3 rd and 4 th grade	P9	Exploring soil profiles in the forest	<ul style="list-style-type: none"> • Solutions • Solution processes 	<ul style="list-style-type: none"> • Investigation • Act autonomously • Interact in heterogeneous groups 	Aims were introduced to the students in a written form at the beginning of the unit	<ul style="list-style-type: none"> • Observation of students' performance • Reflections on performance 	<ul style="list-style-type: none"> • Self-assessment (written) • Peer-assessment (oral) 	<ul style="list-style-type: none"> • Transfer to later situations in the same setting (more teamwork in „Waldlektionen“) 	<ul style="list-style-type: none"> • Teacher evaluation form • Teaching material • Group discussion • Individual interview

A7 continued

Round	School level	Teacher	Criteria related to inquiry units (see section 5.4.3)			Criteria related to formal formative assessment (see section 5.4.4)				Documentation of cases (see section 5.4.2)	
			Inquiry activity	Dimension(s) of openness	assessed competence(s)	Communication of criteria	Data for diagnosis	Formative assessment method(s)	Use of feedback		
1	Upper secondary school; 4 th year	S1	Investigating acoustic noise with several experiments and documenting them	<ul style="list-style-type: none"> • Solutions • Solution processes 	<ul style="list-style-type: none"> • Communication 	Questions that were introduced to the students in a written form before the peer-assessment started	<ul style="list-style-type: none"> • Draft written documentations 	<ul style="list-style-type: none"> • Peer-assessment (written; structured by questions) 	<ul style="list-style-type: none"> • Revision of draft version of documentations 	<ul style="list-style-type: none"> • Teacher evaluation form • Teaching material • Group discussion • Student evaluation form • Individual interview 	
	Upper secondary school; 3 rd year	S2	Conducting individual projects (Mini-Maturaarbeit)	<ul style="list-style-type: none"> • Content • Strategy • Methods • Solution • Solution processes 	<ul style="list-style-type: none"> • Planning • Investigation • Analysis and interpretation • Conclusion and evaluation • Communication • Interact in heterogeneous groups • Act autonomously 	Criteria were introduced to the students in a written form at the beginning of the unit	<ul style="list-style-type: none"> • Observation on students' performance • Discussions on ongoing projects between student groups and teacher • Written reports 	<ul style="list-style-type: none"> • Peer-assessment (written, structured by questions; focussing on transversal competences) • Written teacher assessment (open comments; focussing on domain-specific competences) 	<ul style="list-style-type: none"> • Transfer to "Maturaarbeit" 	<ul style="list-style-type: none"> • Teacher evaluation form • Teaching material • Group discussion • Individual interview 	
	--	S3	--	--	--	--	--	--	--	--	<ul style="list-style-type: none"> • Group discussion
	Upper secondary school; 2 nd year	S4	Investigating pressure and documenting respective experiments	<ul style="list-style-type: none"> • Solutions • Solution processes 	<ul style="list-style-type: none"> • Communication 	Criteria were implicitly clear since they are the same throughout the semester	<ul style="list-style-type: none"> • Draft written reports 	<ul style="list-style-type: none"> • Written teacher assessment (open comments) 	<ul style="list-style-type: none"> • Transfer to very similar task two weeks later (next lab-lesson) 	<ul style="list-style-type: none"> • Teacher evaluation form • Teaching material • Individual interview • Group discussion 	
	Upper secondary school; 1 st year	S5	Constructing a boat & documenting the design/construction process	<ul style="list-style-type: none"> • Strategy • Methods • Solutions • Solution processes 	<ul style="list-style-type: none"> • Planning • Analysis and interpretation • Conclusion and evaluation • Communication • Prediction 	Criteria were introduced to the students at the beginning of the unit	<ul style="list-style-type: none"> • Written reports 	<ul style="list-style-type: none"> • Self-assessment (oral; in groups) 	<ul style="list-style-type: none"> • Transfer to very similar task two weeks later (next lab-lesson) 	<ul style="list-style-type: none"> • Teacher evaluation form • Teaching material • Group discussion 	
	--	S6	--	--	--	--	--	--	--	--	<ul style="list-style-type: none"> • Group discussion
	Upper secondary school; 1 st year	S7	Exploring phenomena related to photosynthesis and cellular respiration	<ul style="list-style-type: none"> • Content • Strategy • Methods • Solution • Solution processes 	<ul style="list-style-type: none"> • Investigation • Communication • Interact in heterogeneous groups • Act autonomously 	Criteria were introduced to the students in a written form at the beginning of the unit	<ul style="list-style-type: none"> • Observation of students' performance 	<ul style="list-style-type: none"> • Peer-assessment (written; with rubric) 	<ul style="list-style-type: none"> • Transfer to later activity 	<ul style="list-style-type: none"> • Teacher evaluation form • Teaching material • Group discussion 	
	--	S8	--	--	--	--	--	--	--	--	<ul style="list-style-type: none"> • Group discussion
	--	S9	--	--	--	--	--	--	--	--	<ul style="list-style-type: none"> • Group discussion
	--	S10	--	--	--	--	--	--	--	--	--
	Upper secondary school; 3 rd year	S11	Investigating ecosystem services	<ul style="list-style-type: none"> • Solutions • Solution processes 	<ul style="list-style-type: none"> • Investigation • Communication 	Criteria were introduced to the students in a written form at the beginning of the unit	<ul style="list-style-type: none"> • Reflections on conduction of investigation • Draft written reports • Presentations 	<ul style="list-style-type: none"> • Self-assessment (written; rubrics and open comments; focussing on investigation) • Peer-assessment (written; rubrics and open comments; focussing on communication) • Written teacher assessment (rubrics and open 	<ul style="list-style-type: none"> • None / unclear 	<ul style="list-style-type: none"> • Teacher evaluation form • Teaching material • Group discussion • Individual interview • Student reflections and student artefacts 	

								comments; focussing on investigation and communication)		
--	--	--	--	--	--	--	--	---	--	--

A7 continued

Round	School level	Teacher	Criteria related to inquiry units (see section 5.4.3)			Criteria related to formal formative assessment (see section 5.4.4)				Documentation of cases (see section 5.4.2)	
			Inquiry activity	Dimension(s) of openness	assessed competence(s)	Communication of criteria	Data for diagnosis	Formative assessment method(s)	Use of feedback		
2	Primary 3 rd grade	P1	Presenting findings on animals	<ul style="list-style-type: none"> • Content • Methods • Solution processes • solutions 	<ul style="list-style-type: none"> • Investigation • Communication • Use tools interactively 	None / unclear	<ul style="list-style-type: none"> • Draft written reports • Draft presentations 	<ul style="list-style-type: none"> • Written teacher assessment (open comments; focussing on investigation and interactive use of tools) • Peer-assessment (oral, focussing on communication) 	<ul style="list-style-type: none"> • Revision of original reports and presentations 	<ul style="list-style-type: none"> • Teacher evaluation form • Teaching material • Group discussion 	
	Primary 5 th grade	P2	Creating crossword puzzle on human body	<ul style="list-style-type: none"> • Content • Solutions 	<ul style="list-style-type: none"> • Investigation • Communication • Use tools interactively 	Criteria were introduced to the students in a written form at the beginning of the unit	<ul style="list-style-type: none"> • Draft crossword puzzle 	<ul style="list-style-type: none"> • Written teacher assessment (open comments; focussing on investigation and on interactive use of tools) • Peer-assessment (oral, focussing on communication) 	<ul style="list-style-type: none"> • Revision of draft crossword puzzles 	<ul style="list-style-type: none"> • Teacher evaluation form • Teaching material • Individual interview • Group discussion 	
	Primary 4 th grade	P3	Presenting findings on animals	<ul style="list-style-type: none"> • Content • Solutions • Solution processes 	<ul style="list-style-type: none"> • Investigation • Communication • Act autonomously 	Criteria were introduced to the students in a written form at the beginning of the unit	<ul style="list-style-type: none"> • Written artefacts 	<ul style="list-style-type: none"> • Written teacher assessment (rubrics and open comments) 	<ul style="list-style-type: none"> • Revision of draft artefacts for domain-specific competences • Transfer to similar situations for transversal competences 	<ul style="list-style-type: none"> • Teacher evaluation form • Teaching material • Group discussion 	
	--	P4	--	--	--	--	--	--	--	--	<ul style="list-style-type: none"> • Group discussion
	Primary 3 rd grade	P5	Observing and documenting the growth of chicks	<ul style="list-style-type: none"> • Solutions • Solution processes 	<ul style="list-style-type: none"> • Investigation • Communication 	Criteria were introduced to the students in a written form at the beginning of the unit	<ul style="list-style-type: none"> • Written descriptions 	<ul style="list-style-type: none"> • Written teacher assessment (open comments) 	<ul style="list-style-type: none"> • Transfer to very similar task the next day (diary on growth of chicks) 	<ul style="list-style-type: none"> • Teacher evaluation form • Teaching material • Observational notes • Student artefacts • Teacher feedback • Individual interview 	
	Primary 4 th grade	P6	Exploring magnetism	<ul style="list-style-type: none"> • Methods • Solutions 	<ul style="list-style-type: none"> • Content knowledge 	None / unclear	<ul style="list-style-type: none"> • Concept maps 	<ul style="list-style-type: none"> • None / unclear 	<ul style="list-style-type: none"> • None / unclear 	<ul style="list-style-type: none"> • Teacher evaluation form • Teaching material • Group discussion 	
	--	P7	--	--	--	--	--	--	--	--	<ul style="list-style-type: none"> • Group discussion
	Primary 3 rd grade	P8	Constructing a bow	<ul style="list-style-type: none"> • Solutions 	<ul style="list-style-type: none"> • Communication • Use tools interactively 	Criteria were provided in the form of a rubric at the beginning of the unit	<ul style="list-style-type: none"> • Reflections on performance 	<ul style="list-style-type: none"> • Self-assessment 	<ul style="list-style-type: none"> • Transfer to subsequent activities 	<ul style="list-style-type: none"> • Teacher evaluation form • Teaching material 	
	Primary 3 rd /4 th grade	P9	Exploring magnetism	---	--	Criteria were introduced to the students in a written form at the beginning of the unit	<ul style="list-style-type: none"> • Written reports 	<ul style="list-style-type: none"> • Written teacher assessment (open comments) 	<ul style="list-style-type: none"> • Transfer to subsequent activities 	<ul style="list-style-type: none"> • Teacher evaluation form • Teaching material • Group discussion • Individual interview 	

A7 continued

Round	School level	Teacher	Criteria related to inquiry units (see section 5.4.3)			Criteria related to formal formative assessment (see section 5.4.4)				Documentation of cases (see section 5.4.2)	
			Inquiry activity	Dimension(s) of openness	assessed competence(s)	Communication of criteria	Data for diagnosis	Formative assessment method(s)	Use of feedback		
2	Upper secondary school; 2 nd year	S1	Using trigonometry in the city	<ul style="list-style-type: none"> • Solution processes 	<ul style="list-style-type: none"> • Communication • Use tools interactively 	Criteria were introduced to the students in a written form at the beginning of the unit	<ul style="list-style-type: none"> • Written reports 	<ul style="list-style-type: none"> • Written teacher assessment (open comments and rubric) 	<ul style="list-style-type: none"> • Transfer to very similar task two weeks later (next lab-lesson) 	<ul style="list-style-type: none"> • Teacher evaluation form • Teaching material • Group discussion • Student artefacts • Teacher feedback 	
	Upper secondary school; 1 st year	S2	Model a mass spectrometer	<ul style="list-style-type: none"> • Methods • Solution processes • Solutions 	<ul style="list-style-type: none"> • Planning • Investigation 	Aims were introduced to the students in a written form at the beginning of the unit	<ul style="list-style-type: none"> • Draft artefact (model of mass spectrometer) 	<ul style="list-style-type: none"> • Written teacher assessment (open comments) 	<ul style="list-style-type: none"> • Revision of draft artefact 	<ul style="list-style-type: none"> • Teacher evaluation form • Teaching material • Group discussion 	
	Upper secondary school; 3 rd year	S3	Investigating separation methods	<ul style="list-style-type: none"> • Solution processes • Solutions 	<ul style="list-style-type: none"> • Communication 	Criteria were elaborated together with students from the peer-assessment	<ul style="list-style-type: none"> • Written lab reports 	<ul style="list-style-type: none"> • Peer-assessment (written; structured by questions) 	<ul style="list-style-type: none"> • Transfer to very similar task two weeks later (next lab-lesson) 	<ul style="list-style-type: none"> • Teacher evaluation form • Teaching material • Group discussion 	
	Upper secondary school; 2 nd year	S4	Investigating electric circuits	<ul style="list-style-type: none"> • Solution processes • Solutions 	<ul style="list-style-type: none"> • Communication 	Criteria were implicitly clear since they are the same throughout the semester	<ul style="list-style-type: none"> • Written lab reports 	<ul style="list-style-type: none"> • Peer-assessment (written; open comments) 	<ul style="list-style-type: none"> • Transfer to very similar task two weeks later (next lab-lesson) 	<ul style="list-style-type: none"> • Teacher evaluation form • Teaching material • Group discussion • Individual interview • Student evaluation form • Student artefacts • Student feedback • Individual interview • Observational notes 	
	Upper secondary school; 1 st year	S5	Deriving rules for the addition of forces	<ul style="list-style-type: none"> • Methods • Solution processes 	<ul style="list-style-type: none"> • Plan • Model • Communication 	Criteria were introduced to the students in a written form at the beginning of the unit	<ul style="list-style-type: none"> • Written reports 	<ul style="list-style-type: none"> • Written teacher assessment (open comments) 	<ul style="list-style-type: none"> • Revision of draft version 	<ul style="list-style-type: none"> • Teacher evaluation form • Teaching material • Group discussion • Observational notes 	
	Upper secondary school; 1 st year	S6	Revising fractions	<ul style="list-style-type: none"> • Solution processes 	<ul style="list-style-type: none"> • Investigation • Act autonomously 	Aims were introduced to the students in a written form at the beginning of the unit	<ul style="list-style-type: none"> • Reflections on the solving processes when working with fractions 	<ul style="list-style-type: none"> • Self-assessment (prompted by questions from teacher during conversation) 	<ul style="list-style-type: none"> • Transfer to subsequent activities 	<ul style="list-style-type: none"> • Teacher evaluation form • Teaching material • Group discussion 	
	Upper secondary school; 2 nd year	S7	Exploring an aquatic ecosystem	<ul style="list-style-type: none"> • Solutions • Solution processes 	<ul style="list-style-type: none"> • Analysis and interpretation • Communication • Use tools interactively 	Criteria were introduced to the students in a written form at the beginning of the unit	<ul style="list-style-type: none"> • Written reports 	<ul style="list-style-type: none"> • written teacher assessment (rubrics and open comments) 	<ul style="list-style-type: none"> • Revision of draft version 	<ul style="list-style-type: none"> • Teacher evaluation form • Teaching material • Group discussion • Student artefacts • Teacher feedback • Individual interview 	
	Upper secondary school; 4 th year	S8	Solving tasks in genetics	<ul style="list-style-type: none"> • None / unclear 	<ul style="list-style-type: none"> • Content-related criteria 	Criteria were introduced to the students in a written form at the beginning of the unit	<ul style="list-style-type: none"> • Draft solutions 	<ul style="list-style-type: none"> • peer-assessment 	<ul style="list-style-type: none"> • Revision of draft solutions 	<ul style="list-style-type: none"> • Teacher evaluation form • Teaching material 	
	--	S9	--	--	--	--	--	--	--	--	<ul style="list-style-type: none"> • Group discussion
	--	S10	--	--	--	--	--	--	--	--	<ul style="list-style-type: none"> • Group discussion
	Upper secondary school; 4 th year	S11	Dissecting a snake	<ul style="list-style-type: none"> • Solution processes 	<ul style="list-style-type: none"> • None / unclear 	None / unclear	<ul style="list-style-type: none"> • None / unclear 	<ul style="list-style-type: none"> • None / unclear 	<ul style="list-style-type: none"> • None / unclear 	<ul style="list-style-type: none"> • None / unclear 	<ul style="list-style-type: none"> • Teacher evaluation form • Teaching material • Group discussion

A7 continued

Round	School level	Teacher	Criteria related to inquiry units (see section 5.4.3)			Criteria related to formal formative assessment (see section 5.4.4)				Documentation of cases (see section 5.4.2)
			Inquiry activity	Dimension(s) of openness	assessed competence(s)	Communication of criteria	Data for diagnosis	Formative assessment method(s)	Use of feedback	
3	--	P1	--	--	--	--	--	--	--	--
	--	P2	--	--	--	--	--	--	--	--
	Primary 5 th grade	P3	Constructing artefacts with explore-it	<ul style="list-style-type: none"> • Content • Methods • Solutions • Solution processes 	<ul style="list-style-type: none"> • Communication • Use tools interactively 	Criteria were provided in the form of a rubric and open questions at the beginning of the unit	<ul style="list-style-type: none"> • Draft presentations 	<ul style="list-style-type: none"> • Peer-assessment (written) 	<ul style="list-style-type: none"> • Revision of draft presentations 	<ul style="list-style-type: none"> • Teacher evaluation form • Teaching material • Group discussion
	--	P4	--	--	--	--	--	--	--	<ul style="list-style-type: none"> • Group discussion
	--	P5	--	--	--	--	--	--	--	--
	Primary 1 st grade	P6	--	--	--	--	--	--	--	<ul style="list-style-type: none"> • Group discussion
	--	P7	--	--	--	--	--	--	--	--
	--	P8	--	--	--	--	--	--	--	<ul style="list-style-type: none"> • Group discussion
	Primary 3 rd / 4 th grade	P9	Growing of beans	<ul style="list-style-type: none"> • Methods • Solutions • Solution processes 	<ul style="list-style-type: none"> • Investigation • Analysis and interpretation • Communication • Use tools interactively 	Criteria implicitly clear	<ul style="list-style-type: none"> • Draft documentation and presentation 	<ul style="list-style-type: none"> • Structured oral comments and written teacher assessment (open comments) 	<ul style="list-style-type: none"> • Transfer to subsequent activities 	<ul style="list-style-type: none"> • Teacher evaluation form • Teaching material • Group discussion
3	Upper secondary school 1st year	S1	Conducting experiments on velocity	<ul style="list-style-type: none"> • Solutions • Solution processes 	<ul style="list-style-type: none"> • Investigation • Analysis and interpretation • Communication 	Criteria were provided in the form of written questions at the beginning of the unit	<ul style="list-style-type: none"> • Reports 	<ul style="list-style-type: none"> • Peer-assessment (written) 	<ul style="list-style-type: none"> • Transfer to very similar task two weeks later (next lab-lesson) 	<ul style="list-style-type: none"> • Teacher evaluation form • Teaching material • Group discussion • Student artefacts • Student feedback
	Upper secondary school 4th year	S2	Conducting matura projects	<ul style="list-style-type: none"> • Content • Strategy • Methods • Solutions • Solution processes 	<ul style="list-style-type: none"> • Orienting and asking questions • Planning • Investigation • Analysis and interpretation • Conclusion and evaluation • Communication • Use tools interactively • Act autonomously 	Criteria were provided in the form of a rubric at the beginning of the unit	<ul style="list-style-type: none"> • Draft matura projects 	<ul style="list-style-type: none"> • Written teacher assessment (rubrics and open comments) 	<ul style="list-style-type: none"> • Transfer to similar activities at university (Semesterarbeiten) 	<ul style="list-style-type: none"> • Teacher evaluation form • Teaching material • Group discussion • Student artefacts • Individual interview
	Upper secondary school 4th grade	S3	Investigating chemical bondings	<ul style="list-style-type: none"> • Solutions • Solution processes 	<ul style="list-style-type: none"> • Communication • Use tools interactively • Act autonomously 	Criteria were provided in the form of written questions at the beginning of the unit	<ul style="list-style-type: none"> • Reflections on presentations 	<ul style="list-style-type: none"> • Self-assessment 	<ul style="list-style-type: none"> • Transfer to subsequent activities 	<ul style="list-style-type: none"> • Teacher evaluation form • Teaching material • Group discussion
	Upper secondary school 2nd year	S4	Investigating electric circuits	<ul style="list-style-type: none"> • Solutions • Solution processes 	<ul style="list-style-type: none"> • Communication • Use tools interactively 	Criteria were provided in a written form at the beginning of the unit	<ul style="list-style-type: none"> • Lab reports 	<ul style="list-style-type: none"> • Peer-assessment (written; open comments) 	<ul style="list-style-type: none"> • Transfer to very similar task two weeks later (next lab-lesson) 	<ul style="list-style-type: none"> • Teacher evaluation form • Teaching material • Student artefacts • Student feedback • Group discussion

A7 continued

Round	School level	Teacher	Criteria related to inquiry units (see section 5.4.3)			Criteria related to formal formative assessment (see section 5.4.4)				Documentation of cases (see section 5.4.2)
			Inquiry activity	Dimension(s) of openness	assessed competence(s)	Communication of criteria	Data for diagnosis	Formative assessment method(s)	Use of feedback	
3	Upper secondary school 2nd year	S5	Elaborating rules to calculate with quadratics	<ul style="list-style-type: none"> Solutions Solution processes 	<ul style="list-style-type: none"> Investigation Documentation 	Criteria were provided in the form of written questions at the beginning of the unit	<ul style="list-style-type: none"> Reports 	<ul style="list-style-type: none"> Self-assessment on investigation Written teacher assessment on documentation 	<ul style="list-style-type: none"> Revision of original report (for investigation) Transfer to very similar task two weeks later (next lab-lesson) for documentation 	<ul style="list-style-type: none"> Teacher evaluation form Teaching material Group discussion
	--	S6	--	--	--	--	--	--	--	<ul style="list-style-type: none"> Group discussion
	Upper secondary school 3rd year	S7	Creating a video visualising genetics	<ul style="list-style-type: none"> Strategy Methods Solutions Solution processes 	<ul style="list-style-type: none"> Model Use tools interactively 	Criteria were provided in a written form at the beginning of the unit	<ul style="list-style-type: none"> videos 	<ul style="list-style-type: none"> Peer-assessment (Written; open comments) 	<ul style="list-style-type: none"> None / unclear 	<ul style="list-style-type: none"> Teacher evaluation form Teaching material Group discussion Individual interview
	Upper secondary school 2nd year	S8	Presenting results of an experiment in a poster	<ul style="list-style-type: none"> Solutions Solution processes 	<ul style="list-style-type: none"> Analysis and interpretation Communication Use tools interactively 	Criteria were introduced to the students in a written form at the beginning of the unit	<ul style="list-style-type: none"> Draft posters 	<ul style="list-style-type: none"> Written teacher assessment 	<ul style="list-style-type: none"> Revision of draft posters 	<ul style="list-style-type: none"> Teacher evaluation form Teaching material Group discussion
	Upper secondary school 1st year	S9	Constructing models of cells	<ul style="list-style-type: none"> Strategy Methods Solution processes solutions 	<ul style="list-style-type: none"> Model 	Criteria were provided in the form of a rubric at the beginning of the unit	<ul style="list-style-type: none"> Models 	<ul style="list-style-type: none"> Peer-assessment (written) 	<ul style="list-style-type: none"> None / unclear 	<ul style="list-style-type: none"> Teacher evaluation form Teaching material Group discussion
	Upper secondary school 3rd year	S10	Elaborating models in electricity	<ul style="list-style-type: none"> Solutions Solution processes 	<ul style="list-style-type: none"> Model Communication 	Criteria were provided in the form of a rubric at the beginning of the unit	<ul style="list-style-type: none"> Draft reports 	<ul style="list-style-type: none"> Peer-assessment (written; rubrics and open comments) 	<ul style="list-style-type: none"> Revision of draft lab reports 	<ul style="list-style-type: none"> Teacher evaluation form Teaching material Group discussion Student evaluation form Student artefacts and peer-assessment Individual interview Observational notes
	Upper secondary school 3rd year	S11	Planning a hedge (ecology)	<ul style="list-style-type: none"> Methods Solutions Solution processes 	<ul style="list-style-type: none"> Communication Use tools interactively 	Criteria were provided in the form of written questions at the beginning of the unit	<ul style="list-style-type: none"> reports 	<ul style="list-style-type: none"> Peer-assessment (written) 	<ul style="list-style-type: none"> Revision of draft report 	<ul style="list-style-type: none"> Teacher evaluation form Teaching material Group discussion

A8. Description of categories for RQ 1

Category	Illustrative quotes
supportive in nature	“formative assessment is meant to support the learning of the students”; “supportive feedback”
providing guidance on next steps in learning for students	“where am I, where do I want to go”; “describes what students already know and also describes what they do not know yet”; “provide guidance”; “coach students in their work”; “hints without giving away the solution”
providing guidance on next steps in teaching for teacher	“I see what is there (preconcepts) and build my teaching on it”; “based on these insights I can tailor my teaching”; “depending on these “checkpoints” I can intervene”
individual and/or part of differentiation	“individual assessment for each student”; “individual progress”; “shows where the individual problems are”; “insight to a person’s actual level of performance”; “individual, so different for different students”; “differentiation dependent on what each student already knows”
prospective rather than retrospective in nature; opposite to summative assessment	“does not involve grading”; “formative assessment is a counterpart to summative assessment”; “it is about the future learning”; “formative assessment is provided before the end of the learning process; so during the learning process”
criterion-based	“teacher observes and takes notes in a rubric on what can be observed”; “criteria are pre-defined”
focussed on a specific set of competences or other learning goals	“Feedback on the personality of the student”, “assessment of the students’ self-regulation skills”; “focussing on social competences”
having an individual reference norm	“in formative assessment, the students is assessed based on individual learning goals”; “assessment based on the individual progress”
grading of the learning process	“assessment which does not focus on the product but on the learning process”
Unclear/ reference to inquiry features	“organise the group, discuss the plans, distribute task amongst group members, documents problem solving process”; “the assessment should provide comparable results to students”
examples of assessment methods	“Written teacher feedback; peer-assessment; self-assessment”

A9. Description of categories for RQ 2

1: Is there a trial?

Did the teacher trial anything at all?

Response format:

Yes

No

2: Is the trial sufficiently documented?

Is the trial documented to extent that it is possible to decide about the subsequent answers on the inquiry- and formative assessment aspects?

Response format:

Yes

No

3: Is the trial inquiry-based?

3a: Student-oriented activity(ies)

Is the unit trialled at least partly student-oriented?

Student-oriented means that the students work relatively independently rather than being guided by the teachers. This could include the students conducting an investigation themselves, document results themselves, discuss the implications of an experiment. Contrary to this, teacher-centred activities include the demonstration of an experiment by the teacher; copy results from the blackboard; answer questions posed by the teacher.

Response format:

Yes

No

3b: Inquiry activities

Which activities were parts of the unit trialled?

Response format:

Orienting and asking questions

Hypothesis generation

Planning

Investigation

Analysis and interpretation

Model

Conclusion and evaluation

Communication

Prediction

Exclusively other activities / unclear / none

3c: Dimensions of openness

In what dimension(s) is the student activity open?

Openness means that not all aspects are pre-defined in a particular task. Instead, the students are able to decide about these aspects themselves.

Response format (multiple responses possible):

Openness in terms of content

Example primary: The topic of the unit is buoyancy. Openness in terms of content means, for example, that students decide themselves whether they want to investigate how quickly different pieces of metal sink, or whether they want to investigate what type of objects sink in water, or ...

Example upper secondary: The mature thesis where students decide themselves what topic they want to work on

Openness in terms of strategy

Example primary: In the above-mentioned context where students investigate how quickly different pieces of metal sink, the students decide themselves whether the different pieces of metal are put into the water at the same time and it is measured what piece reaches the ground first, or whether the pieces are put into the water one after the other and the „time for sinking“ is measured.

Example upper secondary: The investigation is on the differences between serial and parallel circuits. The students decide themselves whether they want to compare the current between the different circuits (quantitatively) or whether they want to compare the brightness of bulbs in the different circuits (qualitatively).

Openness in terms of methods used

Example primary: In the above-mentioned context, the students decide themselves about the size of the box in which they explore buoyancy, they decide themselves about the depth of the water, what pieces of metal will be used, how time will be measured, ...

Example upper secondary: In the above-mentioned context, the students decide themselves which electric source, wire, bulbs etc. to use.

Openness in terms of the number of possible solutions

Example primary: In the above-mentioned context, a possible result could be that a metal ball sinks with the same speed as a metal plate of the same volume. Another possible result could be that a big metal plate sinks more quickly than a small metal plate. And both is correct. (An opposite example: If the task was to find out the temperature or the pH value of the water in the box, only one solution would be correct.

Example upper secondary: Assuming that in the above-mentioned context, an additional task was to document the measurements of the current in a lab journal. There are different possibilities what to write or sketch.

Openness in terms of the number of different solution processes

Example primary: Assuming that the task was to design a model of a car with the materials in the classroom. The students could first produce the axles with the wheels and build the body afterwards or the other way around.

Example upper secondary: In the above-mentioned context with the documentation of an experiment, the students could first measure everything and then write their journals or they could subsequently measure and document.

Unclear / none

3d: Assessed competences

What competence(s) was /were assessed in the trial?

Response format (multiple responses possible):

- Domain-specific competences
 - Orienting and asking questions
 - Hypothesis generation
 - Planning
 - Investigation
 - Analysis and interpretation
 - Model
 - Conclusion and evaluation
 - Communication
 - Prediction
- Transversal competences
 - Use tools interactively including the abilities to use language, symbols and text interactively, to use knowledge and information interactively; and to use technology interactively
 - Interact in heterogeneous groups including the abilities to relate well to others; to co-operate, work in teams and to manage and resolve conflicts
 - Act autonomously including the abilities to act within the big picture; to form and conduct life plans and personal projects; to defend and assert rights, interests, limits and needs
- Exclusively other competences / unclear / none

4: Is there any formative assessment?

4a: Assessment method

What was the formative assessment method trialled?

Response format (fill out separately for every competence assessed):

- Self-assessment (students assess their own work)
- Peer-assessment (students assess their peers' work)
- Written teacher assessment (teacher assesses student work and provides written feedback on it)
- None / other / unclear

4b: Articulation of assessment criteria

Do the students know what criteria their work is assessed by?

Response format:

- Yes, explicitly communicated by the teacher
- Yes, elaborated in the classroom together with the students
- Yes, implicitly clear
- No /unclear

4c: Diagnosis

Is the source of the data for diagnosis (e.g. lab journal, presentation, ...) clear?

Response format:

- Yes
- No

4d: Feedback

Did the students receive any feedback or get to know in any other way what was the result of the diagnosis?

Response format:

- Yes
- No

A9 continued

4e: Use of feedback

Did the students have the opportunity to use the feedback?

Response format (fill out separately for every competence assessed):

- Yes, revision
- Yes, transfer
- No / unclear

4f: Length of cycles

How long was the feedback-cycle; e.g. the time between the reception and the use of feedback?

Response format (fill out separately for every competence assessed):

- Short (minute by minute, day by day)
- Medium (1 to 4 weeks)
- Long (4 weeks to 1 year)
- Unclear

A10. Description of categories for RQs 3.2 and 3.5

Categories	Description	Sub-categories	Examples
Embedding formal formative assessment methods in inquiry-based science education	Advantages and challenges related to when and how to integrate formal formative assessment methods in the semester- or lesson plan	Long-term planning aspects	<ul style="list-style-type: none"> - Peer-assessment activities have to be carefully planned in the course of a semester so that they do not become boring - Peer-assessment is not an assessment method for all students but for social classes mostly
		Shor-term planning aspects	<ul style="list-style-type: none"> - Criteria have to be set in the beginning and cannot be changed or adapted during the course of the unit - Peer-assessment needs students to have their artefacts ready; cannot be postponed to later - Peer-assessment interrupts course of the investigation - Students need different amounts of time to complete feedback to peers
Diagnosis of students' levels of achievement	Advantages and challenges related to the diagnosis of students' levels of achievement	Time pressure	<ul style="list-style-type: none"> - Teacher can take his/her time in deciding about the comments, in setting the priorities in written teacher assessment
		Pre-defined criteria	<ul style="list-style-type: none"> - Diagnosis can also be provided by a team-teaching partner - Diagnosis creates transparency is made in terms of final grading - Peer-assessment can be provided objectively with criteria
		Quality of diagnosis	<ul style="list-style-type: none"> - Diagnosis allows for more nuanced statement than grade - Not all students are equally critical - Students may perceive the peer's level of achievement different than the teacher would
		Individuality	<ul style="list-style-type: none"> - Diagnosis has the potential of setting individual standards

Categories	Description	Sub-categories	Examples
Content of feedback	Advantages and challenges with respect to the content and of the feedback provided	Timing	- Peer-feedback comes immediately
		Focus	- The quantity of the comments is limited; the teachers have to choose what aspects to concentrate on in written teacher assessment - Support minders the student-centred nature of the learning activities
		Quality in terms of content	- Students may have difficulties in distinguishing between sympathy and objective criteria in peer-assessment - If criteria are not clear, students [in peer-assessment] tend to focus on formal issues rather than on more relevant competences - Students are not always honest to themselves in self-assessment
		Quality in terms of language and vocabulary	- Feedback from peers is easily understandable for students because the language and vocabulary used is familiar - Students do not have the vocabulary and style to formulate feedback
		Relation between assessor and assessee	- Feedback, particularly criticism, is easier to accept when it comes from peers - Inhibition level to ask back in case peer-feedback cannot be understood is low
		Potential for enhancement of learning	- Written teacher assessment allows for easily drawing conclusions on the further learning - Written teacher assessment raise the students' awareness of the learning goals
Role of the teacher	Advantages and challenges concerning the responsibilities and the tasks of the teacher in the context of formative assessment	Responsibility for student learning	- Students take responsibility for their own learning - When to interfere if student coaches oversee mistakes?
		Workload for teacher	- Difficult for the teacher to keep the overview over all activities going on - Peer-assessment reduces workload for teacher
		Capacity for individual support	- Peer-assessment provides the teacher with the opportunity to take care of individual difficulties
Use of the feedback by the students	Advantages and challenges related to the question whether and how the students use the feedback they receive	Eagerness of recipients	- Some students may not want any feedback - Engagement with the feedback depends on how critical and eager the individual students are
		Understanding of the feedback	- Student understanding of the feedback may differ from teacher understanding
		Transfer	- Transfer of the feedback to new situations is difficult

Categories	Description	Sub-categories	Examples
Learning effects	Advantages and challenges concerning the learning effects of formative assessment on students	Scientific concepts	<ul style="list-style-type: none"> - Written teacher assessment foster conceptual understanding and content learning - Peer-assessment improves engagement with content
		Nature of science	<ul style="list-style-type: none"> - Peer-assessment fosters students' understanding of why it is important to describe and explain exactly; to write precise protocols; to structure protocols properly; to label sketches
		Science-specific competences	<ul style="list-style-type: none"> - Peer-assessment improves skills / competences assessed - Peer-assessment provides an insight in other students' approaches and solutions which extends personal horizon
		Transversal competences	<ul style="list-style-type: none"> - Peer-assessment fosters collaboration in groups; social development - Peer-assessment fosters communication; feedback culture; distinguish between social effects and subject-specific evaluations - Self-assessment fosters students' abilities to express their opinion and communication skills
		Self-regulated learning	<ul style="list-style-type: none"> - Peer-assessment fosters self-assessment; reflections - Self-assessment fosters students' autonomy as learners
		Other	<ul style="list-style-type: none"> - Assessors can only correct mistakes they already are aware of; they are therefore unable to improve themselves
Social and motivational effects	Advantages and challenges concerning the social and motivational effects of the formative assessment on students	Relation between teacher and student	<ul style="list-style-type: none"> - Written teacher assessment improves the student-teacher relation - Peer-assessment is a way to take students serious and to give value to what they say
		Classroom climate	<ul style="list-style-type: none"> - Peer-assessment is a way for students to show their respect towards other students - Peer-assessment enhances the relation between the students
		motivation	<ul style="list-style-type: none"> - Written teacher assessment show the appreciation of the students' work and therefore motivates students - Peer-assessment is rather boring if all students have the same solution
Documentation	Advantages and challenges that refer to the documentation of diagnosis and feedback in the	Record of feedback for students	<ul style="list-style-type: none"> - Students cannot simply forget the feedback
		Record of feedback for teachers	<ul style="list-style-type: none"> - Teachers need to somewhere take down the hints they have provided to the students

	context of formative assessment	Communication with parents	- The comments are documented which makes them a valuable tool for communication with parents
--	---------------------------------	----------------------------	---

A10 continued

Categories	Description	Sub-categories	Examples
Relation between formative and summative assessment	Advantages and challenges related to the relation between the different components of the assessment system	Relevance of formative assessment	- Peer-assessment is not reliable for summative assessment
		Check-like character	- Assessment hinders the joy and the interest in conducting experiments - Criteria for assessment may hinder the openness of inquiries
Effort needed	Advantages and challenges related to the effort needed so that formative assessment of a high quality results	Time	- Written teacher assessment take a lot of preparation time for the teacher - Self-assessment consumes a lot of lesson time
		Practice	- Peer-assessment does not need a great lot of introduction - Self-assessment needs practice in order to become productive

A11. Description of categories for RQ 3.3

Categories	Description	Examples
Provision of examples of good practice	Teacher quotes suggesting the provision of examples that exemplify the use of formative assessment methods as a means of support	"There should be a collection with good examples of formative assessment activities somewhere, so that not everybody had to re-invent them. These examples could also be adapted or used as inspiration for the own teaching."
Time	Teacher quotes suggesting more time as a means of support	"It just needed more time for preparing appropriate formative assessment activities"
Support from team-teaching partner or another person with a teacher-like function	Teacher quotes suggesting a second adult person in the classroom as a means of support	"It could also be a second teacher, when two classes work in the same project. One of the teachers would basically see that all students work properly and could help providing the materials needed for the project whereas the other teacher could provide individual formative assessment."
Training and coaching to enhance the teachers' assessment literacy	Teacher quotes suggesting the enhancement of teacher assessment literacy as a means of support	"I would like to get a fundus of different methods, an overview of what is out there. [...] I also need some ideas on how to make the formative assessment visible for the parents; I need clear statements so that they know what their children have to work on."
Opportunities and prompts to reflect upon assessment practices	Teacher quotes suggesting the possibility to reflect upon own assessment practices as a means of support	"[...] to become aware of the function of formative assessment; to become aware of the fact that I have been always been doing this but also to learn that there are many more methods and approaches; some of them much more structured and formal than what I have always been doing."
Platform to exchange experiences and problems with peer teachers	Teacher quotes suggesting the opportunity to exchange experiences with peer-teachers as a means of support	"Well, this is not rocket science, but it should be mentioned again: We need more discussion and discourse about formative assessment amongst teachers. We need platforms to discuss our questions, to exchange good ideas."
Clarification of the role of formative assessment and its relation with summative assessment at the level of educational policy	Teacher quotes suggesting the clarification of the term 'formative assessment' or related issues as a means of support	"Formative assessment is really a counterpart to final exams and certificates. So assessment is often focussed on grades. There should be guidelines telling us <the teachers> that not only these grades are important but that assessment that supports learning is also relevant."

A12. Description of categories for RQ 3.6

Categories	Description	Examples
No support needed	Student quote suggesting that no support is needed for peer-assessment.	"I can assess just like that"
Support in formulating feedback	Student quote suggesting that some aid in the formulation of peer-feedback would be supportive.	"I didn't know what comments and suggestions would be constructive in the beginning. So a good example would help."
Structuring questions or criteria to focus	Student quote suggesting that structuring questions or criteria to focus on help for peer-assessment.	"Criteria help to focus on the important parts of a student artefact"
Anonymity	Student quote suggesting that anonymity would be beneficial for peer-assessment.	"Anonymity is important to provide honest feedback"
Access to content knowledge or to the correct solution	Student quote suggesting that access to knowledge is important for good peer-assessment.	"It is important that the assessor has successfully completed the task he/she is assessing."
Exchange with peers or with the teacher	Student quote suggesting that the exchange with peers or with the teacher is important for good peer-assessment.	"It would help if I could talk with the teacher during the peer-assessment, because sometimes, I am not sure about the validity of my feedback"

A13. Description of categories for RQ 4.5

Categories	Description	Examples
Provision of background information / theory	Teacher quotes mentioning the provision of theoretical background information on formative assessment as a helpful element in the study	"Like this, I learn about the background of these methods, I can also look up details again later, [...]"
Provision of concrete methods and examples	Teacher quotes mentioning the information on concrete methods and / or their exemplification in concrete cases as a helpful element in the study	"These examples inspire me, and even though I do not like everything in them, I can still do something similar that goes into the same direction. Many of the examples were for primary and lower secondary level but they can be adapted."
Opportunity to try out methods	Teacher quotes mentioning the prompts and opportunities to try out formative assessment methods as a helpful element of the study	"The chance to try out the methods helps me most. I am just like that. Of course that needs some effort; for preparation and also for the trial itself. I am already thinking about what I could do in the next semester, what would fit the context of a certain topic and class."
Prompts to reflect on own assessment practices	Teacher quotes mentioning the prompts and opportunities to reflect upon their own formative assessment practices as a helpful element in the study	"The project itself was a big help, firstly to become aware of what formative assessment is, and secondly to become aware of my own assessment practices which actually always included some formative assessment even though I did not know the term [...]"
Opportunity to exchange with other teachers	Teacher quotes mentioning the possibility to discuss their formative assessment practices with other teachers as a helpful element in the study	"What is of course useful are the discussions which we have in our meetings. When we are in the small groups with the other physics teachers and start talking, that is really useful. To see, how do the others do in the lab lessons, how do they coach their students, how do they put an assessment method into practice, what unit would be suitable for this."
Broadening of horizon	Teacher quotes mentioning the broadening of their personal teacher horizon as a helpful element in the study	"There are also impulses from the project which I cannot use directly in my teaching, but which broaden my horizon. Like, this could also be done, in a different subject. Or that could be done, at primary school level."