Chapter 6
Students' Perspectives on Peer Assessment

Florence Le Hebel, Costas P. Constantinou, Alena Hospesova, Regula Grob, Monika Holmeier, Pascale Montpied, Marianne Moulin, Jan Petr, Lukáš Rokos, Iva Stuchlíková, Andrée Tiberghien, Olia Tsivitanidou, and Iva Žlábková

Introduction

The role of feedback on student performance is central in formative assessment (Black and Williams 1998) and, obviously, in peer assessment. Peer feedback is expected to support the learning process by providing an intermediate check of the performance according to the criteria and adapted to the individual student, accompanied by comments on strengths, weaknesses, and/or tips for improvement (Falchikov 1996). Learning benefits may arise for students while enacting both the role of peer assessor and peer assessee. Thus, peer assessment can be perceived as a learning tool, since assessing their peers can develop students' judgment-making skills about what constitutes high-quality work and a self-reflection about their own understanding (Topping 2013).

This chapter reports the results of three research studies on peer assessment made in different countries where such practice is unfrequently implemented in as

F. Le Hebel (✉)
Université de Lyon. Laboratoire ICAR, UMR 5191, LLE UMS 3773, Ecole Normale Supérieure de Lyon, Lyon, France
e-mail: florence.le-hebel@ens-lyon.fr

C.P. Constantinou • O. Tsivitanidou
Department of Education, University of Cyprus, Nicosia, Cyprus

A. Hospesova • J. Petr • L. Rokos • I. Stuchlíková • I. Žlábková
Jihočeská univerzita, University of South Bohemia, České Budějovice, Czech Republic

R. Grob • M. Holmeier
School of Education, University of Applied Sciences and Arts Northwestern Switzerland, Basel, Switzerland

P. Montpied • M. Moulin • A. Tiberghien
CNRS, ICAR, UMR 5191, Ecole Normale Supérieure de Lyon, Lyon, France

J. Dolin, R. Evans (eds.), *Transforming Assessment*, Contributions from Science Education Research 4, DOI 10.1007/978-3-319-63248-3_6

a steady classroom organization (France, Switzerland, and the Czech Republic). The three countries are all participating in the EU-funded project ASSIST-ME aiming to develop formative assessment (http://assistme.ku.dk/). The three research studies focus on different competences and different disciplines, but they all involve inquiry-based approaches at primary and secondary school level. The research design is quite similar in the three studies, involving students in a class situation with a teacher who volunteered to participate along with his/her class in the practical implementations. Peer assessment is included in the inquiry-based teaching sequence and students are guaranteed that the research experience would not contribute to their final mark at the end of the semester. In the first step, the students work on tasks individually or in pairs (randomly matched), then their written artefacts are given to another pair (or individual), and they are asked to use rubrics with prespecified criteria for providing feedback to their peers. Once the students complete the peer assessment, they exchange their peer feedback and review it in pairs or individually. Students are allowed to use the peer feedback for revising their artefacts. The students were assured that the study would not contribute to their final mark.

However, as the research questions differ, the data analysis in the three studies varies. In the French study, the data reported in this chapter explore relationships between the success in task processing and the ability to mark a peer's written artefact about the same task. It corresponds in a part of a study investigating to what extent peer assessment helps students to develop understanding and competences involved in the teaching sequence on science. Peer feedback implementation is conducted in physics and geosciences at upper secondary level and focuses on students' investigational competence. In Switzerland, peer feedback is implemented in physics at upper secondary level focusing on modeling competence. Based on a fine-grained analysis of peer feedback comments, the research study examines the type of peer feedback students offer to their peers while assessing their models. In the Czech Republic, the study focuses on students' reflection on peer assessment in inquiry lessons. Peer assessment is conducted in mathematics and biology, at primary and lower secondary level, focusing on problem-solving and investigational competences and is followed by semi-structured interviews with the students. The results of these three studies show some convergent and divergent points of views, and some perspectives on implementing peer feedback as part of formative assessment raised are discussed in the last part of this chapter.

Theoretical Background

The main aim of formative assessment is seen as helping learning. As developed in Chap. 3, the concept of formative assessment is defined differently according to two main conflicting views (Bennett 2011). Some authors consider formative assessment as referring to an instrument and others conceive it as a process. In this project, formative assessment is conceived between these two views. In this chapter, we focus on the process view involving students and teachers (see Fig. 3.1 in Chap. 3, based on Harlen 2013 and modified by Dolin et al.). In particular, we focus on the collection and interpretation of evidences in terms of what it indicates about existing ideas and competences required to meet the lesson goals. In our work, feedback is from student to student. This is consistent with the social constructivist view of learning, which emphasizes the role of interaction in students' understanding construction (see Chap. 3).

Peer assessment is generally an educational arrangement for classmates to judge the level, value, or worth of the products or learning outcomes of their equal-status peers by offering written and/or oral feedback (Topping 2013). The students' picture of themselves or of their "equal-status peers" is one of the crucial points in peer assessment since most often students do not feel fully confident in their own or their peers' knowledge as they are not expert in a subject area. They doubt their peers' ability to assess (Hanrahan and Isaacs 2001; Strijbos et al. 2010; van Gennip et al. 2009; Walker 2001). The peer assessor is usually not regarded as a knowledge authority by an assessee. Students frequently claim that it is the role of the teacher to be the assessor (Brown et al. 2009). Students lack confidence in both their and their peers'abilities as assessors (Ballantyne et al. 2002). Nevertheless, even if the accuracy of peer feedback can vary as students are not experts in the subject area, it can be helpful for learning while students peer assess and when they review their received peer feedback, as they engage in self-reflection processes. In their study, Yang et al. (2006) show that revision following the teacher's feedback is less beneficial for students' understanding than peers' feedback. The authors argue that teacher's feedback was accepted as such, often misinterpreted, and students often considered that no further corrections were expected. whereas peer feedback leads to more discussions and checking for confirmation and consequently a deeper understanding. Moreover, for Strijbos et al. (2010), the qualification "equal-status students" in Topping's definition (1998) might be retained in the sense of class level of students, but there are individual differences that affect perceived status and may impact peer feedback perceptions. Through these research studies, it appears that a student's representation of knowledge authority plays a central role in the dynamics of peer assessment.

Another crucial aspect of peer assessment is the quality of peer feedback. Peer feedback may be delivered either as qualitative (oral or written comments), quantitative (mark), or both. There are various perspectives on peer feedback quality (see Gielen et al. 2010 for an overview). A first perspective defines peer feedback quality as the degree to which a peer's quantitative feedback (mark) matches that assigned by an expert assessor, where scoring validity is the leading concept (Cho et al. 2006; Falchikov and Goldfinch 2000; Van Steendam et al. 2010). A second perspective defines peer feedback quality in terms of content and/or style characteristics. Written comments on a specific piece of work/artefact could vary among peer assessors, because they might focus on different aspects of an assessee's work (Topping 1998). This variability makes the determination of the quality of written comments through specific measures/indices, such as a reliability index, extremely difficult or even impossible in some cases. Some studies try to build a framework for assessing the quality of feedback (Gielen et al. 2010; Prins et al. 2006; Sluijsmans et al. 2002). Different characteristics (summed up in Gielen et al. 2010; Table 1, p. 307) are identified as, for instance, the extent of the peer feedback comments (if the comments are elaborate or superficial) or the justification of peer feedback comments (whether the reasoning underlying the assessors' judgments is revealed to assessees or not).

In this chapter, the peer feedback quality (quantitative and qualitative) is analyzed in the French and the Swiss studies, whereas the Czech study focuses on the students' peer feedback perception.

Research Aims

Different main research questions are addressed:

The French study investigates if a student can actually assess a classmate's work prior to receiving any formal teacher feedback on his own artefact and therefore being in the process of learning from the task through possible exchange with the teacher.

The Swiss study examines the type of peer feedback students generate for their peers while assessing their models in a physics course at upper secondary level.

The Czech research relates to the students' reflection on peer assessment, and more specifically they investigate how students perceive peer feedback they offer and they receive in the context of inquiry lessons.

Relations Between Peer Assessment and Students' Artefact (France)

In France, traditional summative assessment is emphasized; therefore, peer assessment in elementary and secondary school is not a usual practice in most classrooms. Teachers are usually the only ones that provide feedback. However, some French universities initiate and develop students' peer assessment at university (Le Monde 2016).

French official instructions encourage inquiry-based teaching. But this implies to be aware of scientists and students functioning in inquiry. According to Etkina et al. (2010), scientific abilities include but are not limited to collecting and analyzing data from experiments; devising hypotheses and explanations and building theories, assessing, testing, and validating hypotheses and theories, and using specialized ways of representing phenomena and of communicating ideas (Duschl et al. 2007). For scientists, scientific abilities are internalized and become habits of mind to approach new problems. For the students who have not internalized these processes and procedures, scientific abilities are processed that they need to use reflectively and critically (Etkina et al. 2010). Teaching has the necessity to explain and verbalize with students processes and procedures which are internalized by scientists. Inquiry-based activities should allow to make some students' cognitive process visible, not only to the student but also to the whole class in order for them to be a shared knowledge (Tiberghien 2011). Moreover, it leads the teacher to a better understanding of the students and a better adaptation of his/her teaching. In our study, we focus on these three main scientific procedures (from Etkina et al. 2010) that we relate to assessment criteria:

- Make hypotheses/speculations and explanations.
- Collect and analyze data based on experimentations.
- Assess, test, and validate hypotheses.

Peer assessment could help to make these procedures visible to students.

Objectives of the Study

This study aims at investigating to what extent peer assessment helps students to develop understanding and competences involved in the teaching sequence on science. To that end, we use students' written artefacts and videos in the classroom during the activities and peer assessment.

A part of the study, presented in this chapter, investigates if a student can assess a classmate's work successfully even before receiving any formal teacher feedback on his own artefact. The aim here is to explore the relationships between the success in task processing and the ability to assess quantitatively (with a mark) a peer's written artefact about the same task.

More specifically, we worked on two research questions:

1. To what extent are students able to give a mark to each question of their peers' work that is consistent with the mark that an "expert" (teacher/researcher) would have given?
2. To what extent is high achievement of student assessor in the assessed task a mandatory condition to proper marking (consistent with an expert's mark)?

Methodology

We engaged secondary level students (grade 10, 15–16 years old) in a teaching sequence for experimental science. In physics, the sequence was about periodic phenomena (like a heartbeat or beats of the membrane of a loudspeaker, a pendulum, the Earth's movement around the Sun, etc.). In geology, the sequence was about fossil fuels (Table 6.1). In both teaching sequences, two activities were combined with peer assessment (G1 and G2 in geosciences; P1 and P2 in physics).

For each of these activities (G1, G2, P1, and P2), students' pairs had to commit to an investigation-based activity and had to produce a written report of their work. The same pairs of students had to examine afterward the written artefacts of some

Table 6.1 Implementation plan

Activities	Ed. level	Subject	Topic	Number of students
G1 and G2	Grade 10	Geosciences	Fossil energies	168 (6 classes)
P1 and P2	Grade 10	Physics	Periodic phenomena	172 (6 classes)

other group thanks to a structured template that we had previously built. Then, based on the comments they received, they had to revise their initial written work. In all activities, the frame of the investigation-based activity was the same. Students had to:

- Make hypotheses related to a scientific problem
- Use and apply a protocol giving the guidance to design experiments in order to test their hypotheses
- Proceed through the experiment
- Present and analyze the results of their experiment
- Conclude and solve the problem

A major difficulty for students in the case of peer assessment is to determine criteria (OCDE 2005), to see "which goals, which achievements are hidden behind the work they did" (p. 253). To avoid this difficulty, the protocol developed in this experiment does not involve a *collaborative peer assessment scheme* (Stefani 1994, p. 69) where students themselves have to define their own assessment criteria. We chose to provide students with a template giving them several criteria based on the competences they had to use (Examples: Tables 6.2 and 6.3). The templates presented in Tables 6.2 and 6.3 correspond to two activities. In activity G2, the students had to build an analogical model of a real outcrop and a model of an exploitable oil deposit in order to explain why the real outcrop does not contain any oil; in activity P1, the students were supposed to make previsions if three different materials reflect the ultrasound and propose a protocol of an experiment which can check the predictions, carry out the experiment (constrained by the available devices), and compare and discuss the previsions and the results. All templates were built thanks to collaborative work between teachers and researchers.

We observe that when the criteria are close to the question, students understand them better. For instance, in activity P1, the question is "On which objects among three available, do the ultasounds reflect?" The criterion "the predictions are justified" (fully, incomplete, totally unrelevant) is better understood than "the justification is drawn on elements of knowledge other than ultrasounds."

For each criterion, peer assessors had to give a mark (four-level scale: one is the lowest level of achievement and four is the highest) and a written comment (justification) in order to justify and help assessees to improve their artefact. In geosciences, criteria relate to complex competences (as presented in the French curricula) such as the ability to make empirical observations, interpret them, and conclude (criterion 3). In physics, teachers and researchers worked on the elicitation of these complex competences in order to provide students with criteria which assess rather specific components of the answer (Table 6.3). For instance, in assessing if the pro-

Table 6.2 Assessment template (Geosciences: Activity G2)

Criteria	Description	Mark (1–4)	Justification
1	Schematic drawings are well accomplished and fit with the instructions of the methodology sheet		
2	The model allows us to understand the situation described in the document: The links between elements of reality and elements of the model are effectively made; the oil movement is shown (e.g.,with an arrow)		
3	The observations made on the model are well expressed and interpreted		

Table 6.3 Assessment template (Physics—Activity P1)

Criteria	Description	Mark (1–4)	Justification
1	A prediction is made and justified		
2	The protocol is relevant to test the prediction		
3	The results of the experiment are clearly presented and appear coherent		
4	The experimental results are interpreted in relation to the predictions, and conclusions are drawn with regard to the questions on the reflection of ultrasounds by different materials		

tocol is relevant for testing the hypothesis (criterion 2), peer assessors have to check the coherence of the hypotheses and the protocol and not the quality of the hypotheses or of the protocol.

We focus here on the students' ability to offer the quantitative feedback (i.e., providing a mark) consistent with the peer assesses' level of achievement as determined by an expert. We gathered data in 12 10-grade classes (6 in physics, 6 in geosciences, Table 6.4). We collected written artefacts of each group. Our methodology of analysis consists of cross-examining, with the help of the ASSIST-ME Project coding tool (ASSIST-ME report D5.11 http://assistme.ku.dk) written assessment with the work previously done by the students. We first compared the peer assessors' marking with an expert's marking (Q1) to see if students are able to give a proper mark to their classmates. Then we combined this information with the peer assessors' level to determine if a high level of achievement during the activity is a precondition for peer assessment (Q2).

Data Analysis

The aim of our analysis is to study the students' ability to assess other students' artefacts by providing a mark on a four-level scale. Peer assessors had to provide a mark for each assessed criterion. Table 6.4 gives the number of marks expected for

Table 6.4 Data collected for each activity (all criteria)

Activity	Topic	Number of expected marks	Number of marks collected
G1	The origin of fossil fuels	2 per group Total: 180 (90 groups)	170 (94%)
G2	The formation of a fuel deposit	3 per group Total: 228 (76 groups)	228 (100%)
P1	Periodic phenomena 1	4 per group (4 classes) 5 per group (2 classes) Total: 358 (83 groups)	358 (100%)
P2	Periodic phenomena 2	5 per group (4 classes) 6 per group (2 classes) Total: 389 (72 groups)	360 (93%)

each group (which is also the number of criteria in the template), the number of marks effectively collected for all those criteria (column 4).

Our analysis relies on a comparison between the marks given by an expert (teacher and/or researcher) with the marks given by the students. We studied for all activities and classes the gap between the expert's mark and the student's mark for the same criterion. With a four-level scale for grading, we can get seven values for this gap (from −3 to 3):

- When the gap is null, expert and assessors gave the same mark.
- When the gap is positive, the expert gave a higher mark than the assessors.
- When the gap is negative, the assessors gave a higher mark than the expert.

Regarding our first research question (Q1), we considered that the mark is consistent with the level of the assessees when the gap has a value of {−1; 0; 1} and is inconsistent when the gap has a value of {−3; −2; 2; 3}. Then, we cross-examine this information with the level of achievement of the assessors wondering if a high level of achievement is a precondition for peer assessment (Q2). In order to do this, we have counted how many students:

- Succeeded in the initial task (Mark of their own work given by the expert ≥3) and succeeded in the marking (Gap ≤1)
- Succeeded in the initial task (Mark ≥3) and failed in the marking (Gap ≥2)
- Failed in the initial task (Mark ≤2) and succeeded in the marking (Gap ≤1)
- Failed in the initial task (Mark ≤2) and failed in the marking (Gap ≥2)

Results

In most of the 12 classes, students were committed to the peer assessment task even if initially some of them expressed doubts or concerns in their abilities to be good and fair assessors. For example, students sometimes asked their teachers how they could possibly evaluate the work of their peers without knowing if their own answer was correct or without having successfully accomplished the activity, which shows their willingness to do well by their peers.

Table 6.5 Gaps between assessors' and experts' marks (Note that the number of marks can vary slightly from those in Table 6.3 because we did not take into account the anonymous artefacts)

	G1		**G2**		**GeoS**		**P1**		**P2**		**Physics**		**Total**	
	#	**%**	**#**	**%**	**#**	**%**	**#**	**%**	**#**	**%**	**#**	**%**	**#**	**%**
Number of marks	161		221		382		326		286		612		994	
Gap = 0	87	54	108	49	195	51	143	44	132	46	275	45	470	47
Gap = 1	10	6	14	6	24	6	56	17	29	10	85	14	109	11
Gap = -1	48	30	60	27	108	28	93	29	70	24	163	27	271	27
Total Consistent						85 %						86 %	850	85 %
Gap = 2	1	1	8	4	9	2	6	2	14	5	20	3	29	3
Gap = -2	8	5	15	7	23	6	25	8	31	11	56	9	79	8
Gap = 3	2	1	9	4	11	3	1	0	3	1	4	1	15	2
Gap = -3	5	3	7	3	12	3	2	1	7	2	9	1	21	2
Total Incon-sistent						15 %						14 %	144	15 %

The analyses of the gap between the experts' and the students' marks (Table 6.5) show that, for the most part (85%), students are able to provide a mark consistent with the assessor' levels:

- More than 47% of the students' marks exactly equal the expert mark showing that the assessors gave the proper mark to their classmates (470 out of 994).
- More than 85% of the gaps (850 out of 994) valued between {−1, 0, 1} showing a marking consistent with the assessor level.
- Only 15% of the gaps showed inconsistency (108 with {−2; 2}, 36 with {−3; 3}).

When not equal to 0, the gaps are for the most part negative, meaning that students gave a better mark than the experts. This tendency (which occurred in all classes except one in geosciences for G2[1] and one in physics for P1[2]) could be explained in two ways:

- The assessors wanted to be nice to their classmates and friends (stated by some students in the interviews).
- The assessors didn't recognize a mistake or a lack in their peers' written artefact.

[1] In this specific class, 21 gaps were positive and 21 negative.

[2] In this specific class, 32 gaps were positive and 21 negative.

Table 6.6 Relations between success/failure on a criterion in the activity and the success/failure in the assessment (for all criteria listed by activity)

Initial activity	Marking	G1 N= 156	G2 N=207	P1 N=333	P2 N=327	Total N=1023
Success **N = 719**	Success	126 (90%)	142 (92%)	192 (93%)	194 (87%)	654 (91%)
	Failure	13 (10%)	12 (8%)	13 (7%)	27 (13%)	65 (9%)
Failure **N = 304**	Success	16 (94%)	38 (71%)	107 (83%)	77 (72%)	238 (78%)
	Failure	1 (6%)	15 (29%)	21 (17%)	29 (28%)	66 (22%)

As a preliminary work regarding our second research question, we have numbered, for all criteria listed by activvity, the number of times that a two-student pair have succeeded/failed in the activity and the number of times that a student pair have succeeded in giving/failing to give a proper mark. In total, 752 expert's marks show a success regarding a criterion during the initial activity and 898 students' marks were consistent with the assessors' level. This means that some students who didn't reach a high level of achievement during the activity on a specific criterion (mark ≤2) were able to provide a proper mark on this specific criterion (gap ≤1 with the expert mark).

More specifically, each time we could, we have numbered the times where a two-student pair succeeded/failed in the activity and succeeded/failed in the marking on the same criterion (Table 6.6). We can see (column7) that there are 65 times where students who succeeded in the activity (on a specific criterion) didn't give a proper mark on the same criterion. Conversely, there are 238 times where students who didn't reach a high level of achievement managed to give a proper mark to their peers on the same criterion.

The percentages in the last column of Table 6.6 show that a majority of students were able to invalidate/validate the work of their peers (Q1) even if they did not succeed in the task (Q2). Indeed, 78% of the students with a low achievement on a criterion manage to give a proper mark on the same criterion. This can seem surprising, but the fact that we provided students with meaningful criteria with some details during peer assessment may have helped some students to reconsider and maybe to understand the objectives of the activity. In the first activity in geosciences, we can see that this rate is even better. In 16 cases out of 17 (column 1), students were able to give a proper mark even if they didn't manage to reach the expectation for the same criterion during the activity. This can be explained by the fact that this activity was easier, in terms of knowledge and competences involved, than G2. Students had to make an observation, draw a sample containing information, and use documents in order to discuss the origin of fossil fuels. The criteria were related to the ability to communicate their observation scientifically through the drawing (which students are used to doing) and to write an argumentative paragraph explaining the origin of fossil fuels based on written documents. In G2, they had to build an analogical model of a real situation (complex task) in order to understand how fuel deposits are formed and explain why the situation under study does not contain oil anymore. Criteria (Table 6.1) are related to the ability to establish those links and use those links in an argumentative paragraph. The same goes in physics; P2's activity and assessment were more difficult than P1's.

Conclusion

In this work, we investigate student's abilities to mark a peer's written artefact in relation to the given criteria and in regard to their own artefact in the same activity. We show that most of the students, without any teacher correction, are able to give a proper mark, consistent with the one that an expert could give. Even part of the students with a low achievement rate manage to mark their peers. But one of our major concerns is to know if the fact that most students have the capacity to give a proper mark also implies that those students understand the tasks and their own possible mistakes. The analysis of the comments students gave each other does not show a deeper student understanding. The following study investigates further student comments. A first analysis of the geosciences activity, based on video data of students doing their activity and assessing their peers, tends to show that assessing their peers does not help students to be aware of their own misunderstanding (Le Hebel et al. 2016). The authors conclude that there is the necessity of a crucial phase of discussion after assessment between peers and with the teacher during the phase of correction.

Another concern is to characterize the type of tasks which are most suitable for peer assessment. It seems that one of the main features for this suitability is that students are asked to assess the process (even if they do not assess all the steps of the process) rather than factual knowledge (they may lack). As stated above, it seems that peer assessment, despite the information provided in the criteria, does not always help students to recognize their own mistakes and improve their understanding and knowledge. When assessors and assessees made the same mistake, there is no possibility for them to recognize their misunderstanding even with the help of the criteria. Sometimes, even students who have succeeded in the task are not aware of a misunderstanding in their peers' artefact. These observations show the importance of the role that teachers need to play during peer assessment. We think that a time of common correction by the teacher with the whole class is necessary at the end of the peer assessment. Moreover, peer assessment templates are a way for teachers to share explicit assessment criteria and give students guidance on what is expected from them. Collaborative work and discussion of these criteria (during the peer assessment time and/or during the correction of the activity or possibly during collaborative construction of peer assessment templates) can enhance the awareness of their students' needs in order to improve their knowledge and work processes. By the way, we built new templates for peer assessment in geosciences including an answer along with the associated assessment criteria, each one being located on the answer and a grid with formulated justifications making a scale to assess the level of achievement for each criterion. This grid is a way to permit student discussion and reflection on the answers of their peers. It is a source for students who felt unconfident with the previous template to be certain of their assessment. The fact that the answer is given is maybe not in alignment with the idea that peer assessment is a way for students to improve by themselves. In our case, students have to correct a functional schema which necessitates the whole process understanding. The interpretation cannot be done element by element. The qualitative study that we are currently running, based on comments and video analysis, shows that the improvement is not that valuable even if students seem to be able to give a mark. We think that peer assessment must be a time for metacognition and a reflection on the expectations. This type of grid, considered as support for this metacognitive work, helps students to establish links between the answer, the criteria, the objectives, and the knowledge involved in the activity, even if in some cases it will not be sufficient.

As a perspective, we also want to emphasize the fact that it seems necessary that the responsibility of assessing other students' artefacts given by the teacher to the students should be in accordance with the usual responsibilities given by the teacher. Taking more responsibility in the validation of their artefacts is a main way for students to develop their competences and autonomy but not only for this type of activity, which is, as Allal (1999) underlines, in the spirit of investigation activities. In consideration, teachers have to work on the formulation and moreover the transmission of assessment criteria (OCDE, 2005) in coherence with the classroom practices that they instigate. Students must be aware of what is expected of them.

Examining the Peer Feedback that Secondary School Students Generate for Their Peers' Models in Science (Switzerland)

The educational system in Switzerland allows for much individuality at the level of school units (Husfeld 2009) as well as at the level of the individual teacher at all school levels (Kronig 2009). This is particularly true for the culture of assessment (Vögeli-Mantovani 1999). The use of formative assessment strategies including peer assessment therefore depends largely on the initiative of the school or the initiative of the individual teacher (Smit 2009). In particular, at upper secondary school level, the external stimuli from curricula and from textbooks play a minor role, and the interest in formative assessment is almost exclusively triggered by ongoing activities at the compulsory school levels.

Educational research indicates that the development of modeling competence is facilitating students learning of science, about science, and of how to do science (Saari and Viiri 2003; Schwarz and White 2005). The modeling competence could be fostered in the context of modeling-based learning, which refers to "learning through construction and refinement of scientific models by students" (Nicolaou and Constantinou 2014; p. 55). In consideration of the complexity of acquiring and mastering the model construction competence itself, it might be even more demanding to request secondary school students to offer feedback to their peers, after having evaluated their peers' models. Model critiquing involves engaging students in discussing the quality of models for evaluation and revision (Chang et al. 2010; Schwarz and White 2005; Schwarz et al. 2009). Critiquing is an important scientific practice that needs to be addressed in science classrooms (Duschl et al. 2007), and it could be practiced through peer assessment activities. A few studies (e.g., Chang and Chang 2013; Pluta et al. 2011) have provided evidence specific to the educational value of teaching-learning activities that involve the critiquing of models. Chang and Chang (2013) stressed the need for more research in this direction, with the ultimate goal being the identification of what students can do when assessing peers' models (Chang and Chang 2013). Considering the lack of previous research on what student critiquing of peers' models entails, we sought to further examine this issue in this study.

Objectives and Research Questions

The present study focuses on peer assessment of the model construction component. It aims at investigating the peer feedback that secondary school students generate for their peers' model in science.

More specifically, two research questions are addressed:

1. What are the characteristics of the qualitative peer feedback that secondary school students generate while assessing their peers' models in the context of light and color in a physics course?
2. What is the relation of qualitative (written comments) and quantitative (ratings in the rubric) peer feedback provided by secondary school students?

Methodology

Participants

The physics teacher involved in this study volunteered to participate with his class in a classroom implementation of formative assessment methods developed by the ASSIST-ME project. The typical instructional format of the classroom was characterized by student group activities in the laboratory due to the nature of the course (physics) in combination with lectures given by the teacher. The learning goals of the course entailed the development of conceptual understanding of physical phenomena and the development of experimental and problem-solving skills. In the meetings that took place among the teacher and the researchers involved in this study, for organizational purposes, it was clarified by the teacher that the peer assessment method was not a commonly practiced method in his class, except the cases in which students exchange oral peer feedback among them in a nonformal setting.

The class comprised 22 students of the 11th grade at a Gymnasium (i.e., the highest track at upper secondary school) in Switzerland. The students were assured that the study would not contribute to their final mark. They worked in randomized pairs throughout most parts of the intervention, and the pairs did not change during the intervention. There were 11 groups of 2 students. However, the students worked individually when enacting the peer assessor role.

Teaching-Learning Sequence

The sequence was grounded in collaborative modeling-based learning. The students worked through learning material on the topic colors and light. The curriculum material required the students to work with a list of hands-on experiments on additive and subtractive color mixing (McDermott et al. 1996). After completing the experiments, students were instructed to draw inferences relying on their observations and the gathered data. Their inferences were explicitly expected to lead to a scientific model which represents, interprets, and predicts the additive and subtractive color mixing of light. In order to do this, students were provided with a sheet of paper, color pencils, and a list of specifications that they were asked to consider when developing their model. The list of specifications constituted three benchmarks that students should consider while developing their model which are the following: the model should (i) represent, (ii) interpret, and (iii) predict the additive and subtractive color mixing of light. The list of specifications was in line with the assessment criteria that were given later on during the peer assessment activity to students. Overall, it took the student groups five lessons of 45 minutes to complete this sequence.

The Process of Peer Assessment

As soon as the students had finalized their models in their home groups, they exchanged them with models of other groups; i.e., two groups assessed their models mutually. The exchange pairs were randomly defined by the teacher. Peer assessors used a rating scale with eight prespecified assessment criteria (e.g., Does the model appropriately represent what subtractive color mixing is? Is it explained and justified in the model, which are the primary colors? Can the formation of white or black be derived from the model?), which were in line with the list of specifications that was given to students while constructing their models. Assessors rated their peers' models on all criteria according to a four-point Likert scale (i.e., (1) unsatisfactory; (2) moderately satisfactory; (3) good; (4) (fully) satisfactory/excellent). A fifth column was provided next to the rating scale for each criterion for the provision of written comments. Along with ratings, assessors were prompted by the teacher to provide written feedback (for each criterion separately) to assessee groups, in which they were to explain the reasoning behind their ratings and provide judgments and suggestions for revisions (Table 6.7).

The students were instructed to individually assess the model of the peer groups assigned to them. On average, it took each peer assessor 15 minutes to complete the

Table 6.7 The four-point rating scale with eight prespecified assessment criteria. This is an example of a fulfilled rating scale provided by student coded as 8A from the assessor group 8 to the assessee group 9. The model of group 9 is presented in Fig. 6.1. Note: The text was translated from German to English

Assessment criteria for the model for color mixing: Assess your peers' model according to the following criteria. Provide a meaningful comment in the space provided.

Assessee Group: 9		Your code: 8A				
		1 = unsatisfactory, 2 = moderately satisfactory, 3 = good, 4 = (fully) satisfactory/excellent				
	Assessment criteria	1	2	3	4	Your comments
1	Does the model appropriately represent the primary colors?			x		Only with drawing, without explanation. Bottom right and the light circles you can recognize them. But it is not explained
2	Does the model appropriately represent what additive color mixing is?			x		Very good model, maybe a small text to explain … again no description of what is what
3	Does the model appropriately represent what subtractive color mixing is?			x	x	Illustration provided
4	Does the model appropriately represent how color filters work?				x	Illustration provided
5	Does the model explain how color filters work?		x			Description missing
6	Is it explained and justified in the model, which are the primary colors?		x			Only with drawing, without explanation. Cyan and magenta were not different from yellow without blue
7	Can the formation of white or black color be derived from the model?		x			White is to be recognized in the color circle on the bottom right. Black is illustrated at the top as white. However, it is not explained how black is formed, and it is not said that black is formed
8	Can you predict what is formed in the additive mixture of yellow and cyan, relying on your peers' model?			x		It does not say what color is formed. However, good representation. The color circle shows a lot

assessment. Once the students had completed the assessment of their fellow students' models, they exchanged their peer feedback and reviewed it in collaboration with their groupmate, deciding whether to make any revisions to their model.

Data Collection

At the beginning of the intervention, a consent form was signed by the students' parents, allowing us to use the collected data anonymously and for research purposes. The filled-out rating scales with peer feedback comments produced by students, along with their schoolwork, own constructed models (e.g., in Table 6.7 and Fig. 6.1) were collected to allow us to address our research objective.

Data Analysis

We used a mixed-method approach that involved both qualitative and quantitative analyses of the data. In particular, data were firstly analyzed qualitatively and then also treated quantitatively with the use of the SPSS™.

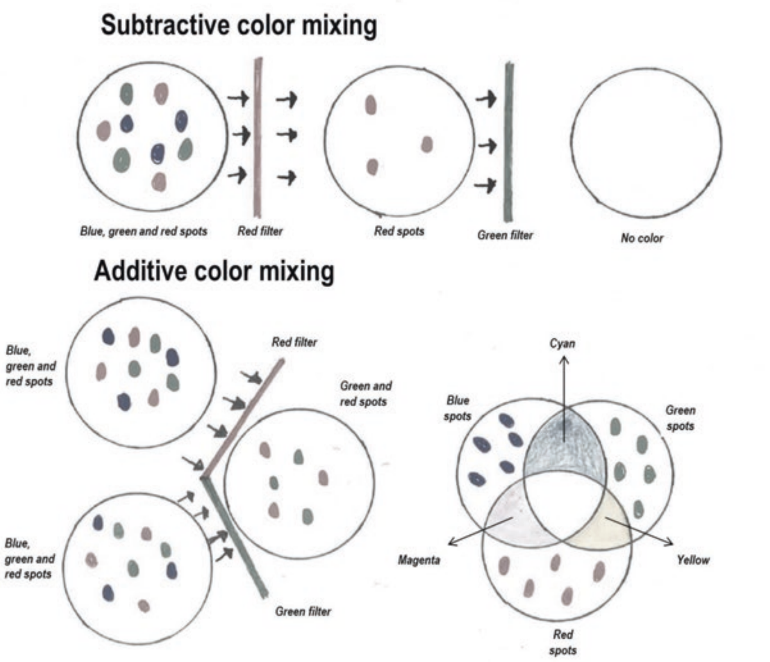
Determining the level of quality of peer feedback requires examining the quality of both the quantitative and qualitative feedback. In this study, we mainly focus on the qualitative part of peer feedback, that is, to say the written comments provided in the four-Likert scale rubric along with the ratings, because it has been emphasized in prior research that qualitative feedback is more important than quantitative (e.g., Topping et al. 2000).

To analyze peer feedback written comments, we developed and further used a coding scheme including the following dimensions:

1. Comprehensiveness (whether the assessors drew on the intended assessment criteria and to what extent).
2. Validity of peer feedback comments (their scientific accuracy and their correspondence to the models assessed).
3. Verification of comments (peer feedback comments were perceived as "positive" when including references to what the assessees had already achieved with respect to the list of specifications; likewise peer feedback comments were perceived as "negative" when including references to what the assessees had yet to achieve with respect to the list of specifications).
4. Justification of positive and negative comments (i.e., justification offered by the assessor(s) on what the assessees had achieved or not yet achieved with respect to the list of specifications related to the modeling competence).
5. guidance provided by the assessor(s) to the assessee(s) on how to proceed with possible revisions. We perceived "guidance" as statements which could potentially help the assessee(s) to improve their models.

The peer feedback comments of each student were coded separately. Each complete sentence included in the peer feedback comments, from each student, was analyzed with respect to all the aforementioned categories (comprehensiveness, validity, verification, justification, guidance). The resulting codes were further used quantitatively, for running nonparametric correlations (Kendall's T_b) between the coded written comments and the ratings assigned by students to each criterion of the rating scale.

A possible internal consistency between the qualitative (comments) and quantitative (ratings) peer feedback provided can be used as an indicator of the quality of peer feedback (Hovardas et al. 2014). For that reason, we further examined whether there is a statistically significant correlation between the quantitative score assigned and the number of references to what the assessees have already achieved and what



-All three primary colors → white light
-No color → black
-How a filter functions: a blue filter filters everything but blue
-Likewise for a cyan filter → green, blue

- ~~3~~ Grundfarben → Weiss
- Ohne Farbe → schwarz
- ~~Filter fü~~ ~~F~~ Funktion Filter:
Blau filtert alles ausser Blau
Cyan → Grün Blau

Subtractive color mix:
The first filter filters the white light and allows only some colours to pass. Because the green filter only allows green light to pass, while only red light falls on the filter, nothing comes through.
→<u>This is how black is formed</u>

This can be adopted for all colors.

- subtraktive Farbmischung
Der erste Filter filtert das weisse Licht
und lässt nur rotes Licht durch.
Da der Grünfilter nur grünes
Licht durchlässt und nur rotes
Licht auf den Filter trifft, ~~bleibt~~
~~keine~~ kann nichts durchdringen.
=> Schwarz

Dies kann man auf alle Farben
anwenden.

Fig. 6.1 This is an example of the type of models that students constructed and the kind of revisions applied by assessees after receiving peer feedback. The initial model of Group 9 included three representations (*on the top*). The students added in their revised model an explanatory text for their representations (*on the bottom*). Note: The text was translated from German to English (*bold font letters*). Also, labels have been added for the each color for readability purposes (*letters in italics*)

the assessees have not yet achieved in respect to the modeling competence (i.e., if quantitative scores are positively correlated with number of positive judgments references to what the assessees have already achieved in respect to the modeling competences and negatively correlated with number of references to what the assessees have not yet achieved in respect to the modeling competence).To estimate the interrater reliability of the data coding, a second coder who had not participated in the first round of coding repeated the coding process for 40% of the peer feedback data and the students' models. Each peer feedback comment (one complete sentence) provided by each assessor for each assessment criterion of the rating scale that students used while giving feedback was rated for 16 items of the coding scheme, addressing the aforementioned dimensions. In all cases, the two raters involved in the data analysis process were also asked to justify their reasoning for their ratings and/or provide illustrative examples. Krippendorff's alpha was calculated above 0.79 for the coding of all data. The differences in the assigned codes were resolved through discussion.

Results

Type of Qualitative Peer Feedback Provided

Assessors provided feedback comments which were found to carry affective connotations. In particular, peer assessors provided on average 3 and in total 75 comments which carried affective connotations (illustrative quotes: "Well done, well described," "Very nice illustration, I like the models"). Assessors tended to provide a balance of statements which are likely to serve as discouraging and encouraging feedback for the assessees.

In terms of the comprehensiveness of peer feedback comments, the analysis revealed that assessors took into account most of the assessment criteria which were given to them while peer assessing their peers' models (i.e., they rated and commented on average on 6.54 criteria out of the total of 8 criteria of the given rubric) and they drew on them in a rather thorough manner. In particular, in the criteria where the assessors did not provide full marks in their rating with the four-point Likert scale, they justified the awarded low marks by suggesting what was missing and what could be added in their peers' models; in other words, they were specifying what their peers did not manage to achieve in respect to their modeling competence. With respect to the validity of peer feedback comments, the findings have shown that assessors' judgments with respect to the criteria which they attended to were mostly valid. In a few cases, students provided invalid comments to their peers, and those comments were found to be related to misconceptions identified in their own models.

Peer feedback comments were found to be critical enough, as the assessors tended to include in their feedback comments more negative and fewer positive comments. In particular, the data analysis revealed that assessors provided on average 3.71 positive comments (M = 3.71; total = 89; SD = 2.87; illustrative quote:

"Graphically very well recognizable and easily comprehensible") and on average 5.58 negative comments (M = 5.58; total = 134; SD = 2.28; illustrative quote: "...but it could be added why it is called additive color mixing."). Positive and negative peer feedback comments were mostly justified. Assessors provided on average 3.0 and in total 73 justified positive feedback comments (SD =2.76) (illustrative quote: "Good, because it is explained and graphically illustrated.") and on average 0.67 and in total 16 positive comments (SD =1.09) which were not justified (illustrative quote: "Your model looks great."). Similarly, the assessors provided on average 5.33 and in total 128 negative comments (SD =2.28) that were justified (e.g., "...the thing with white containing all colors not enough. Below maybe use red, blue and green dots instead of white ones.") and on average 0.25 and in total 6 negative comments (SD =0.53) which were not justified ("Wrong.").

Moreover, assessors provided the assessees with guidance statements (average = five comments per assessor) on what the assessees needed to further achieve to improve the quality of their models. Overall, all the assessors provided 115 comments which could be perceived as guidance statements. Five statements were not related to the competence of interest but with superficial aspects related with the appearance of the model (e.g., legibility of writing; illustrative quote*:* "The writing is not always legible."), whereas the rest were related (e.g., "An interpretational text is missing, so you should add this in your model."). In addition, specific guidance about next concrete steps, provided by the assessors, was found to be present to a very small extent (e.g., "I understand from your model how white is formed but not black.") up to a great extent (e.g., "The primary colors are green, red and blue. Orange & violet are mixed colors as well as magenta, cyan. Therefore, you should revise your model according to this."). The guidance statements provided by the assessors were found to be mostly valid; this means that the students, as assessors, were able not only to identify most of the weaknesses in their peers' models but also to guide them on the next steps that were to be taken to improve their models.

Relation of Qualitative (Comments) and Quantitative (Ratings in the Rubric) Peer Feedback Provided

As part of the quality check of peer feedback, we ran Kendall's T_b correlations between the written feedback given per criterion on what the assessees have achieved or not in respect to the modeling competence and the score assigned to the corresponding criterion, to check whether there was internal consistency between the quantitative and qualitative aspects of the provided peer feedback. The analysis revealed that the mean scores, which were assigned by the assessors in the four-point Likert scale rubric along with their comments in the criteria which they attended, were negatively correlated with the number of negative comments (i.e., references to what the assessees had not yet achieved in respect to the modeling competence) (Kendall's $T_b = -0.749$; $p < 0.01$). This means that the lower the score an assessor was assigning to a certain criterion, the more the possibilities were for providing written comments to what the assessees had not yet achieved. No statistically significant correlation was found between the mean of scores and the number of references to what the assessees had already achieved in respect to the modeling competence.

Discussion and Conclusion

This study focused on examining the type of peer feedback that secondary school students generate while assessing their peers' models in the topic of light and color in a physics course. The findings of this study have shown that the participants were committed to their assessor role in a satisfactory way. They offered justified positive and negative comments, as well as guidance statements to their peers, acknowledging what assessees had already achieved or not yet achieved in respect to the list of specifications which was given to them and was related to the modeling competence. Assessees were also capable of offering guidelines to their peer for the next steps to be taken for improving their models. To follow on from this, it was found that assessors tended to provide a balance of statements which were likely to serve as discouraging and encouraging feedback for the assessees. This indicates that assessors could identify flaws and shortcomings in assessees' models (thus offering negative comments), justify their comments, and provide suggestions for improvements (offering guidance). Suggestions and recommendations for possible ways for improvement (Hovardas et al. 2014; Strijbos and Sluijsmans 2010), as well as justified comments (Gielen et al. 2010; Narciss 2008; Narciss and Huth 2006), have been identified by researchers as essential characteristics of constructive peer feedback. This commitment to the assessor role was also confirmed by the study's consistency checks, which revealed that the peer assessors were attentive while assessing. For instance, it was found that scores given by peer assessors were negatively correlated with number of references to what the assessees had not yet achieved in respect to the modeling competence, which indicates that quantitative aspects of peer feedback (i.e., negative ratings) were consistent with qualitative aspects of peer feedback (i.e., references to what the assessees had not yet achieved in respect to the modeling competence). Also the findings of this study have revealed that peer assessors drew on almost all available criteria of the rubric while providing feedback. Lastly, peer feedback comments were found to be mostly scientifically accurate and consistent throughout, even though in a few cases assessors offered invalid peer feedback comments which were related to misconceptions identified in their own models. Overall, the findings suggest that students have the beginnings of providing peer feedback of good quality in modeling-based learning.

The findings of this study have some practical implications. The fact that secondary school students were committed to the peer assessor role renders peer assessment as a promising formative assessment method and learning tool in modeling-based learning. Students, as assessors, were found to be capable of providing their peers with constructive feedback in a satisfactory manner, which could be ultimately used to foster the learning progress of both. Teachers could use the peer feedback as a probe to students' model construction practices and understanding of scientific concepts.

However, considering that few students—who addressed some of the intended specifications in an invalid way in their own model—provided their peers with invalid feedback comments with respect to those specifications, it is implied that the validity of peer feedback comments is related to students' understanding of the topic under emphasis. Previous studies have shown that the quality of peer feedback is associated with students' understanding of the subject studied (Ballantyne et al. 2002). Peer assessment requires students to use not only their assessment skills (Sluijsmans 2002) but also their knowledge on the content/topic in order to review, clarify, and correct peers' work. Teachers who engage their students in peer assessment activities in modeling-based teaching should safeguard that students, as assessees, do not receive wrong signals from their peers through the peer feedback comments, especially when the assessors have not completely comprehended what constitutes a good scientific model or how the phenomenon operates. This requires support mechanisms for the teachers themselves, via tools or guidelines on how to filter peer feedback content, before exchanging it among students or on how to easily identify students' models of low validity and henceforth filter or closely moderate the feedback comments that those students deliver to their peers.

This study was carried out in the context of a regular physics course, thus offering ecological validity to the aforementioned findings. On the other hand, this study reveals limitations related to the sample size, which was small (22 students). Future studies should replicate the findings of the present study with bigger sample sizes. The conclusions of this study should be limited to the particular characteristics and affordances of the students involved.

Overall peer assessment in modeling-based learning constitutes an area that calls for future research to further explore the potential benefits that peer assessment may entail in this context.

Student Perspective on Introduction of Formative Peer Assessment (Czech Republic)

The idea of formative assessment which emerged in the 1990s was only slowly introduced in the Czech educational context (c. f. Žlábkova and Rokos 2013). The Czech Framework Educational Programme for Basic Education implemented in 2007 aims at gradually accomplishing changes in the assessment of pupils toward diagnostics on an ongoing basis, assessment of pupils' achievement history, and a wider use of verbal assessment (compared to marks). The practice of assessment ought to be driven by guidelines embedded in each school rules document. This document should describe principles and methods of assessment and self-assessment of learning outcomes and conduct of students, including the acquisition of data for evaluation and criteria for the evaluation. Self-assessment is thus explicitly stated as a necessary part of the school assessment practice, but explicit peer assessment is not. Peer assessment is seen rather as a particular method in cooperative teaching (Novotná and Krabsová 2013).

In recent years, there has been great interest in upscaling the formative assessment as there is a need for change in learning culture (Education at a Glance 2015: OECD Indicators Czech Republic 2015; Santiago et al. 2012). The problem is that though there are some examples of good practice (c. f. Košťálová et al. 2008; Kratochvílová 2011; Slavík 2003), they are not empirically studied and focus mostly on selected subjects and educational levels. Empirical research which would provide some evidence on the effectiveness of using various formative assessment methods (FAM) is quite scarce (Novotná and Krabsová 2013).

Some types of formative assessment are seen as more or less embedded in common Czech teaching culture, like teacher provided on-the-fly assessment or to a lesser degree written feedback (Laufková and Novotná 2014; Lukášová 2012; Košťálová et al. 2008; Novotná and Krabsová 2013), and methodological literature for teachers (e.g., Kratochvílová 2011; Starý 2006) pays more attention to these forms of formative assessment. In these materials, formative peer assessment is mentioned only as a supplementary option (e.g., Košťálová et al. 2008), probably also due to the fact that peer assessment as a form of classroom communication is not very frequent (Šeďová et al. 2012) and formative peer assessment was not yet empirically studied in Czech schools. The aim of this study is to investigate students' views on introduction of formative peer assessment in inquiry-based lessons of primary mathematics and biology and secondary biology.

Objectives of the Study and Research Questions

The main research questions relate to the students' reflection of peer assessment in inquiry lessons aiming at development of problem-solving and empirical investigation competencies. The central issues under investigation are:

- How do students perceive formative and summative peer assessment?
- Do they prefer peer or teacher assessment?
- How do they reason out their preferences?
- Which difficulties do students experience in providing their peers with assessment?
- How did they perceive the peer assessment that they received and its value for their learning?

Methodology

Formative peer assessment was investigated in three samples of students in different subjects (primary mathematics, primary biology, and secondary biology), who tried to provide their classmates with oral (second graders) or written (third to ninth

Table 6.8 Implementation plan

Subject	Education level	Topic	Organization of work and peer assessment	Number of students
Primary mathematics	Grades 2, 4, 5	Basic geometrical shapes and their area, big numbers	Group discussion (2nd grade), pairs or small groups	113 (6 classes)
Primary biology	Grades 3, 4	Germination	Individual	79 (experimental groups in 6 classes)
Secondary biology	Grades 6,7 8, 9	Germination human physiology	Individual	76 (experimental groups in 6 classes)

graders) peer assessment while solving inquiry tasks oriented toward problem-solving (PS) and empirical investigation (EI) competence development (Table 6.8). Formative peer assessment was structured (by forms in written form and by teachers questions in oral form) and was, as such, a new method of assessment for both students and teachers. Also the inquiry tasks were developed ad hoc for this study (see description below).

Experimental teaching units consisted of one to six inquiry tasks of different complexity, which usually took 2 months of teaching. In biology classes, we also use bogus peer feedback (the written feedback was in fact provided by a teacher, but the students did not know this) to see whether the students react more to the authorship or the content of the assessment. Inquiry task in mathematics usually presented natural life problem, and the students were asked to search for possible solution, describe the procedure, and use mathematical notation of it. The task in biology asked for finding factors which influence particular phenomena (germination, breathing frequency) and develop and experiment which could validate hypothesized factors.

Whereas students participating in the primary mathematics group worked mostly in small groups and provided peer feedback also for groups of mates (in group discussion in the second grade, in rubrics with oral comments in the fourth grade), the students participating in both primary and lower secondary biology groups worked on tasks individually and used only rubrics for providing their peers with the feedback (Table 6.9). In these rubrics the assessors provided formative feedback (e.g., assessment of experiment design, of the possibility to collect the necessary evidence and use this evidence for testing the stated hypothesis) and also summative feedback, using a mark (1–5) summarizing the overall assessment (hence, the name summative feedback) of the task solution and a justification of the mark.

Structured interviews contained a set of questions to prompt students' reflection of their experience (questions relevant to the peer assessment are in Table 6.9). The interviews were transcribed. Where appropriate (questions 4–7), we also used data from the students who only received peer feedback but did not provide it (e.g., due to taking longer on their own solution, or as a member of the control group; primary biology classes N = 61; lower secondary biology classes N = 67).

Table 6.9 Structured interview questions

Questions	Template codes
1. Have you been assessing the work of some of your peers?	
2. Do you think that you did well in assessing the peer's work?	(1) Yes or mostly yes, (2) no or mostly no, (3) I do not know or other answer
3. Did you have any difficulties when assessing the work of your peers? What kind and why?	(1) Lack of knowledge or skills (related to task, criteria, proper formulation)[a], (2) social regards (positive and negative)[b], (3) would also rather know other solutions[c], (4) bad handwriting—not sure whether understood correctly, (5) no difficulties reported
4. Would you rather get the feedback on your work from your peers or from the teacher? Why so?	(1) From teachers, (2) from peer, (3) does not matter
5. Do you think that there are any differences between the teachers' and peers' assessment of your work? What sort?	
6. When you received the feedback from the peer, did it help you to improve your solution?	(1) Yes or mostly yes, (2) no or mostly no, (3) I do not know or other answer
7. When you received the feedback, what information interested you the most?	(1) Mark, (2) comments to your work, (3) both

Examples of coded utterances:

[a] Firstly, we need to know what it is about and what should be done, and I am not sure what is correct; I did not know how to describe what the flower needs, the assessment itself, whether it is correct or not; I did not understand the picture, it was difficult

[b] It is difficult if you know that it is your friend; I did not want to assess him badly

[c] I did not know whether it was correct, I may have solved it in another way, I would rather see more solutions

The answers to questions were coded by three coders according to the same template. The differences in codes were discussed until consensus was reached.

Data Analysis

We used deductive thematic coding with a template approach (Crabtree and Miller 1999). An a priori template of coding categories in the form of a codebook was applied as a means of organizing the data for subsequent interpretation. When using a template, a researcher defines the template (or codebook) before commencing an in-depth analysis of the data. The codebook is sometimes based on a preliminary scan of the text, but for this study, the template was developed a priori, based on the research question. The codebook presented coding categories (template codes in Table 6.8) and examples of codes from utterances of different age cohorts. A

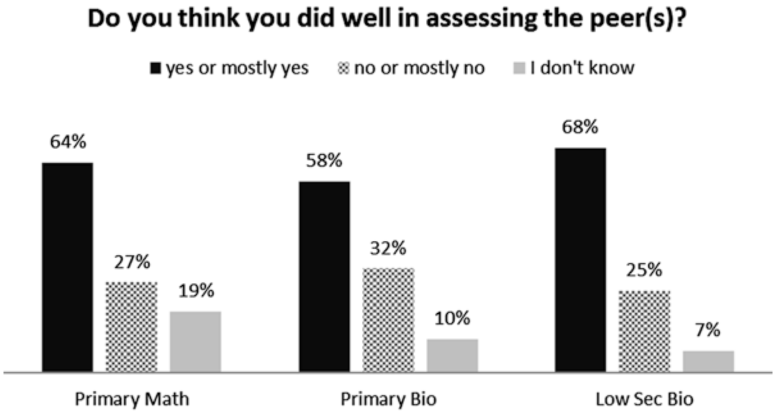


Fig. 6.2 Subjective perception of doing well in assessing the peers

quantitative survey of received codes was then compiled and compared between the age and subject cohorts.

Results

The participating students did not have any previous experience with peer assessment, and therefore we wanted to know how well they did in providing the feedback on peers' work.

The students felt relatively competent in assessing their peers, regardless of their age and regardless of the subject or form of feedback (see Fig. 6.2).

Nonetheless, the students reported some difficulties they encountered when providing the feedback. Figure 6.3 shows the most frequent areas of difficulty that the students mentioned.

The most frequent and most consistently reported difficulty relates to the lack of knowledge or skills necessary for correct assessment, which was associated either with uncertainty about the solution of the inquiry task or the criteria for the assessment. For example, the task on breathing frequency asked the students to develop an experiment which could decide whether there is a relation between intensity of motion and breathing frequency. The assessor was prompted to assess design of the experiment, whether it will provide data on investigated relation and whether the experiment could be realized and how many relevant factors are included. The students who experienced uncertainty about the solution or criteria mentioned that, e.g., "it was difficult to decide whether it is correct or not," "I did it differently and I am not sure that this could be realized," etc. Dealing with this uncertainty in the classroom is crucial for implementation of peer feedback in a broader context. What is further evident is that the primary mathematics groups, which mostly worked in small teams, was more sensitive to social regards when providing the feedback (e.g., worries of negative emotions of peers, bad own feelings when assessing the bad

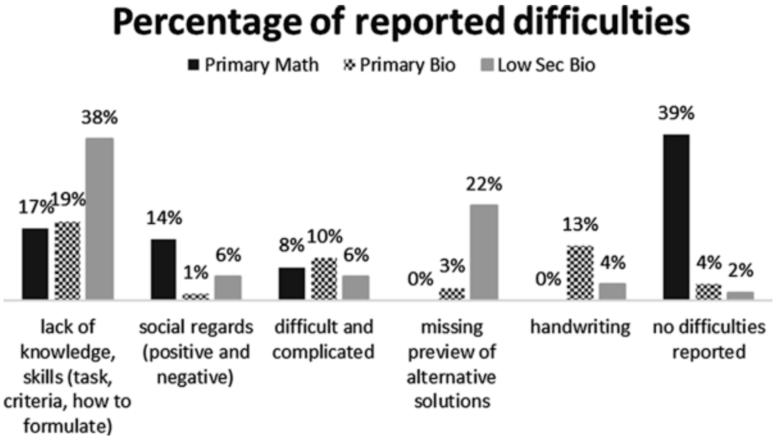


Fig. 6.3 Percentage of reported difficulties when providing peers with feedback

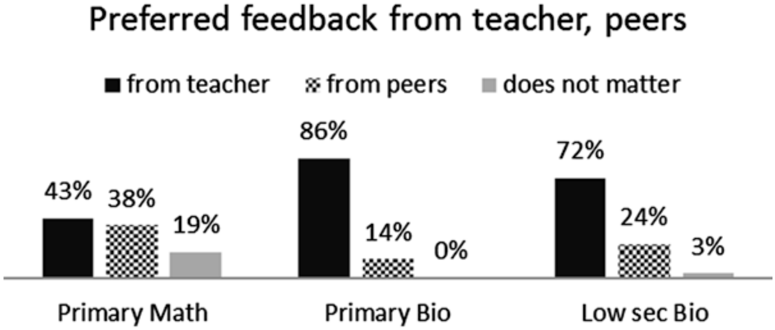


Fig. 6.4 Preferences of teacher feedback and peer feedback

artefact of a good friend, etc.). On the contrary, students from the primary mathematics group did not mention the need to see alternative solutions or trouble with handwriting, and more than one third of this group did not report any difficulties.

An important issue emerging from the interviews with students (and teachers) was the preference of feedback either from the teacher or from peers. Before the start of experimental teaching, the teachers frequently expected that the fact that peer feedback is expressed in more accessible peer language could be a potential advantage of implementation of formative peer assessment. We therefore asked the students about their preferences and reasons for them.

Again, there is a difference between biology and mathematics groups, students in the mathematics group do not show such a pronounced prevalence of teacher assessment preference as the other two cohorts do (Fig. 6.4). The mathematics group students valued the ideas of their peer assessors as easily accessible and did not worry that much about their own mistakes when these are mentioned by peers, unlike when it is the teacher. Biology group students significantly preferred the

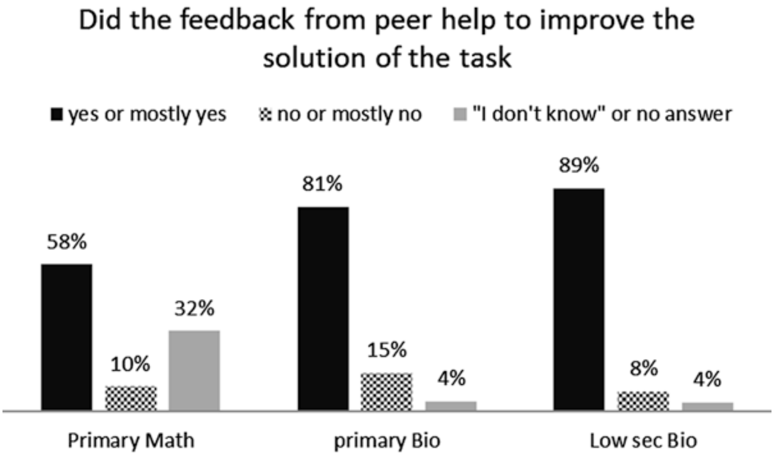


Fig. 6.5 Students' perception of formative impact of peer feedback

feedback to come from the teacher and came to the conclusion it was more reliable and effective (pointing out weak or bad parts of the solution more directly).

What is considered as essential for formative assessment is the informative part of feedback which can foster further learning. For the question whether the feedback from peers was seen as valuable for further improvement of the artefact/solution, the codes were summarized into three categories: positive statement, negative statement, and uncertain standpoint (see Fig. 6.5). In the secondary biology group, we used also bogus peer feedback where the feedback written in a rubric was in fact provided by a teacher. We wanted to know whether the quality and particularity of feedback when thought to be provided by peers is used more for the improvement of their own work. No differences were found; some students did not pay attention to the utilization of any feedback, regardless of its quality. What is even more important is the fact that students reported that the feedback was valuable for improving their solution, but did not actually use it, as could be seen from the working sheets.

We asked all the students, who worked in biology groups where comments as well as marks were provided by assessors in the feedback rubric, what they were more interested in.

In spite of the fact that the students first searched for the mark, which was also frequently mentioned by teachers, students reported that they were more interested in comments that they received on their work (see Fig. 6.6).

Discussion

A student's perspective on their own learning is an important resource of valuable information for teachers, especially when they are trying to implement new approaches to teaching. To understand the ways in which their practice influences student learning, they need to listen to students' accounts of their learning

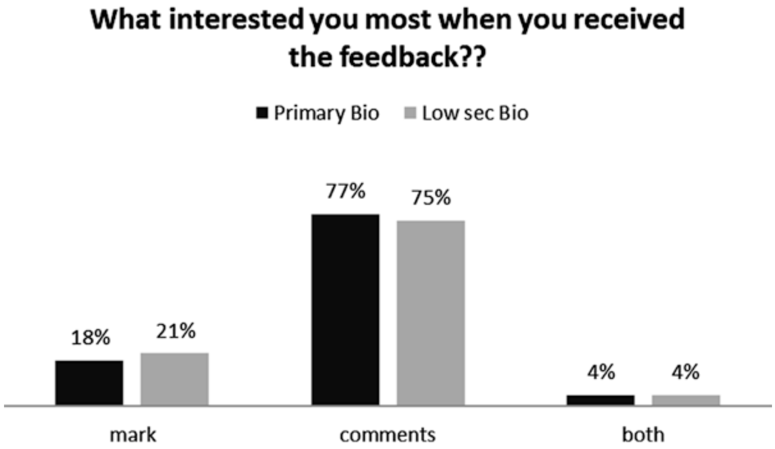


Fig. 6.6 Marks and/or verbal assessment preference

experiences (Kane and Chimwayange 2014). Students participating in this study experienced peer assessment in inquiry lessons for the first time. The results show that the students considered themselves as rather capable of providing peer assessment. Nonetheless, they also experienced some difficulties. The most pronounced category of these was a subjectively perceived lack of domain knowledge needed for evaluating correctness of the peer's solution or procedural knowledge related to providing proper hints or advice for peers on how to proceed further. Students also expressed that they lacked the opportunity to see other alternative solutions which made them uncertain in assessing the peer's work. It seems that specification of proper criteria in rubrics is not enough and that students need some support in applying them, at least when they are not yet used to it.

Results also showed that students also take into account the social context of peer assessment. It was more visible in our primary mathematics sample, where the students worked in small groups. Working in groups can make the students more aware of the socio-emotional context of assessment. Working in small teams led, on the contrary, to more opportunities to discuss the solution and the formulation of proper feedback before giving it to the other team. These students therefore did not report interest in other solutions. The mathematics group students also reported a very small proportion of trouble with handwriting and a high percentage of no difficulty. This could also be influenced by the fact that students in these groups frequently used the possibility to add a verbal explanation to what is written when problems in understanding arise or provided feedback within group discussion (second grade) instead of in writing.

Students in biology groups, who were working and providing peer assessment individually, preferred a teacher's assessment over that of their peers significantly more than the mathematics group. It could be related to the fact that biology groups worked on more convergent inquiry tasks than the mathematics groups where the tasks were more divergent and sharing them in small groups usually led to broader discussion in the whole class. The use of rubrics in our biology groups may also have reduced the advantage of "peer language" as it was more structured, and the proportion of open expression in the feedback was smaller. It seems that a lack of opportunity to broadly discuss the solution and the criteria for its assessment increases students' reservations about plausibility, correctness, and completeness of the feedback from peers.

Conclusions

The students' interviews provided important information on the main challenges and hindrances the students faced in providing peer assessment. As this method of formative assessment is not common in Czech learning culture, students experienced peer assessment for the first time. The students seemed to prefer feedback from a teacher over the assessment of their peers because they see it as more reliable and relevant. But this preference is dependent on the organization of students' work and peer assessment. When students worked in small groups, where they had more opportunities to discuss the solution, this prevalence disappeared. A similar finding is that the work in small groups was accompanied by fewer difficulties encountered in providing the feedback. Discussions in small groups may alleviate the uncertainty associated with a lack of factual knowledge and with any socio-emotional consequences of assessment. Though the students reported that the feedback helped them to improve the solution of the task, they carried out these improvements only in small numbers (which was found in the analyses of revised protocols). This could be due to a lack of time, motivation, or fatigue of students (mathematics lessons videodata provide evidence that it was the case in two inquiry tasks). It is an important message that teachers must pay attention to.

Students also commented on the difficulties they had in assessing inquiry of their peers (uncertainty about correctness, about a considerate way of reporting mistakes, etc.) and advantages they saw in peer feedback over feedback from the teacher, although these issues are probably dependent on the organizational forms of instructions (e.g., individual vs. small group work). Students mostly appreciated the inquiry tasks and would like to extend the proportion of such learning to their everyday lessons. They perceived peer feedback as a new alternative to assessment, but having experienced it for the first time, they remained more in favor of feedback provided by a teacher.

Synthesis of Results and Perspectives

In the French study, results show most of the students are able to give a quantitative feedback with a proper mark, consistent with one that an expert (researcher/teacher) could give without any teacher correction. In this study related to investigation competence, even students with a low achievement rate manage to mark their peers. However, the fact that students with a low achievement rate manage to mark their peers may indicate at least a surface understanding (e.g., if the question and the answer respect drawing codes, the wording of the answer uses similar words to the question), even if the link between the quality of students' peer feedback and the quality of their own artefacts may mean that justification and comments reveal a deeper conceptual understanding. The findings of the Swiss study show that students are able to offer their peers justified negative and positive comments as well as guidelines for the next steps to be taken, acknowledging what the assessees have already achieved or not in respect to the modeling competence. Quantitative aspects of peer feedback (i.e., negative rating) are consistent with qualitative aspects of peer feedback (i.e., references to what the assessees have not yet achieved in respect to the modeling competence). In the Czech study, one of their conclusions is that students perceived peer feedback as a new alternative to assessment; however, following their first experience of it, they express their preference for the feedback provided by a teacher. But this preference is dependent on the organization of students' work and peer assessment. When students worked in small groups, this prevalence disappeared. The authors suggest that the kind of task (e.g., in biology or mathematics) could also have an impact on students' feedback perception.

The three studies conclude the necessity of allowing the sharing of "knowledge authority" in the classroom to evolve. It needs to be integrated in usual classroom practice. However, researchers have a divergent view on the sharing of responsibility for validation of knowledge between the student and the teacher. For instance, the authors of the Swiss study propose that to prevent assessees from receiving wrong signals from their peers through the peer feedback comments, especially when the assessors have not completely comprehended what constitutes a good scientific model or how the phenomenon operates, the teachers themselves might be able via appropriate tools and guidelines to filter feedback content before exchanging it among students. The French authors have a different position on this point and propose to give assessors more autonomy in assessing the other students' artefact. That is in line with previous studies showing that peer feedback leads to more discussions and checking for confirmation and consequently a deeper understanding than teacher feedback, accepted as such and often misinterpreted (e.g., Yang et al. 2006). For the French authors, the phase of discussion after assessment between peers is crucial, and the phase of institutionalization by the teacher during the phase of correction is also essential. Students need to be aware of what the teacher and their peers expect from them as assessees and assessors. Moreover, peer assessment templates could be a way for teachers to share explicit assessment criteria and give students guidance on what is expected from them. Collaborative work and discussion of these criteria (during the peer assessment time and/or during the correction of the activity or possibly during collaborative construction of peer assessment templates) could enhance the awareness of their students' needs in order to improve their knowledge. Peer assessment could be a way to trigger metacognitive work on knowledge and competences in science.

Moreover, in the countries participating in ASSIST-ME, most of the teachers agree on the usefulness and effectiveness of formative assessment, but they all express lack of time to implement it in classroom (see Chap. 3). In these three studies, included in this chapter, teachers all express that peer assessment is time-consuming. As it is already pointed out in Chap. 3, we think it is crucial that teachers do not conceive peer assessment as an add-on to the usual teaching but perceive it as central and integrated part of teaching. It implies that teachers implement peer assessment in a continuity with the other assessments (formative and summative). For instance, they have to use the same criteria in their different formative and summative assessments.

Further research from these three studies could focus on to what extent peer assessment helps students to develop understanding and competences involved in the teaching sequence for science and how to characterize the type of tasks that are most suitable for peer assessment.

References

Allal, L. (1999). *Impliquer l'apprenant dans le processus d'évaluation: promesses et pièges de l'autoévaluation*. In Depover, C., & Noël, B. (1999). L'évaluation des compétences et des processus cognitifs. Modèles, pratiques et contextes (pp. 35–56).

Ballantyne, R., Hughes, K., & Mylonas, A. (2002). Developing procedures for implementing peer assessment in large classes using an action research process. *Assessment & Evaluation in Higher Education, 27*, 427–441.

Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy & Practice, 18*(1), 5–25.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy and Practice*, *5*, 7e74.

Brown, G. T., Irving, S. E., Peterson, E. R., & Hirschfeld, G. H. (2009). Use of interactive–informal assessment practices: New Zealand secondary students' conceptions of assessment. *Learning and Instruction, 19*(2), 97–111.

Chang, H. Y., & Chang, H. C. (2013). Scaffolding students' online critiquing of expert-and peer-generated molecular models of chemical reactions. *International Journal of Science Education, 35*(12), 2028–2056.

Chang, H. Y., Quintana, C., & Krajcik, J. S. (2010). The impact of designing and evaluating molecular animations on how well middle school students understand the particulate nature of matter. *Science Education, 94*(1), 73–94.

Cho, K., Schunn, C. D., & Charney, D. (2006). Commenting on writing: Typology and perceived helpfulness of comments from novice peer reviewers and subject matter experts. *Written Communication, 23*(3), 260–294.

Crabtree, B. F., & Miller, W. L. (Eds.). (1999). *Doing qualitative research* (pp. 163–177). Sage Publications.

Duschl, R., Schweingruber, H. A., & Shouse, A. W. (Eds.). (2007). *Taking science to school: Learning and teaching science in grades K-8*. Washington, DC: National Academies Press.

Education at a Glance. (2015). OECD Indicators Czech Republic. DOI:10.1787/eag-2015-51

Etkina, E., Karelina, A., Ruibal-Villasenor, M., Rosengrant, D., Jordan, R., & Hmelo-Silver, C. E. (2010). Design and reflection help students develop scientific abilities: Learning in introductory physics laboratories. *The Journal of the Learning Sciences, 19*(1), 54–98.

Falchikov, N. (1996). Improving learning through critical peer feedback and reflection. In *Different Approaches: Theory and Practice in Higher Education. Proceedings of HERDSA Conference*.

Falchikov, N., & Goldfinch, J. (2000). Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. *Review of Educational Research, 70*(3), 287–322.

Gielen, S., Peeters, E., Dochy, F., Onghena, P., & Struyven, K. (2010). Improving the effectiveness of peer feedback for learning. *Learning and Instruction, 20*(4), 304–315.

Hanrahan, S. J., & Isaacs, G. (2001). Assessing self-and peer-assessment: The students' views. *Higher Education Research & Development, 20*(1), 53–70.

Harlen, W. (2013). *Assessment & inquiry-based science education: Issues in policy and practice*. Trieste: Global Network of Science Academies.

Hovardas, T., Tsivitanidou, O. E., & Zacharias, C. Z. (2014). Peer versus Expert feedback: Investigating the quality of peer feedback among secondary school students assessing each other's science web-portfolios. *Computers & Education, 71*, 133–152.

Husfeld, V. (2009). Aus der Praxis der Leistungsbeurteilung. In D. Fischer, A. Strittmatter, & U. Vögeli-Mantovani (Eds.), *Noten, was denn sonst? Leistungsbeurteilung und -bewertung* (pp. 33–40). Zürich: Verlag LCH.

Kane, R. G., & Chimwayange, C. (2014). Teacher action research and student voice: Making sense of learning in secondary school. *Action Research, 12*(1), 52–77.

Košťálová, H., Miková, Š., & Stang, J. (2008). *Školní hodnocení žáků a studentů se zaměřením na slovní hodnocení* [School assessment of students]. Praha: Portál.

Kratochvílová, J. (2011). *Systém hodnocení a sebehodnocení žáků.* [System of assessment and self-assessment of pupils]. Brno: MSD.

Kronig, W. (2009). Schulnoten–Glasperlen des Bildungssystems. *Noten, was denn sonst*, 27–31.

Laufková, V., & Novotná, K. (2014). *Školní hodnocení z pohledu žáků* [School assessment from the students' perspective] *Orbis scholae*, (*1)*, 111–127.

Le Hebel, F., Montpied, P., & Moulin, M. (2016). Does peer assessment help students' understanding? Paper presented at the International Conference of East-Asian Association for Science Education, Tokyo, Japan.

Le Monde. (2016). *Quand les étudiants se forment entre eux.* http://www.lemonde.fr/campus/article/2016/03/09/quand-les-etudiants-se-forment-entre-eux_4878964_4401467.html

Lukášová, H. (2012). *Proměny pojetí vzdělávání a školního hodnocení* [Changes in conceptualisation of the education and school assessment]. Praha: Asociace waldorfských škol ČR.

McDermott, L. C., Shaffer, P. S., & Rosenquist, M. L. The Physics Education Group.(1996). *Physics by inquiry*.

Narciss, S. (2008). Feedback strategies for interactive learning tasks. In J. M. Spector, M. D. Merrill, J. J. G. Van Merrie¨nboer, & M. P. Driscoll (Eds.), *Handbook of research on educational communications and technology* (3rd ed., pp. 125–143). Mahwah: Erlbaum.

Narciss, S., & Huth, K. (2006). Fostering achievement and motivation with bug–related tutoring feedback in a computer–based training for written subtraction. *Learning and Instruction, 16*, 310–322.

Nicolaou, C. T., & Constantinou, C. P. (2014). Assessment of the modeling competence: A systematic review and synthesis of empirical research. *Educational Research Review, 13*, 52–73.

Novotná, K., & Krabsová, V. (2013). Formativní hodnocení: Případová studie [Formative assessment: case study]. *Pedagogika, 63*(3), 355–371.

Organisation de coopération et de développement économique. (2005). *L'évaluation formative: pour un meilleur apprentissage dans les classes secondaires*. Paris: OCDE.

Pluta, W. J., Chinn, C. A., & Duncan, R. G. (2011). Learners' epistemic criteria for good scientific models. *Journal of Research in Science Teaching, 48*(5), 486–511.

Prins, F. J., Sluijsmans, D. M., & Kirschner, P. A. (2006). Feedback for general practitioners in training: Quality, styles, and preferences. *Advances in Health Sciences Education, 11*(3), 289–303.

Saari, H., & Viiri, J. (2003). A research-based teaching sequence for teaching the concept of modelling to seventh-grade students. *International Journal of Science Education, 25*(11), 1333–1352.

Santiago, P., Gilmore, A., Nusche, D., & Sammons, P. (2012). *OECD Reviews of Evaluation and Assessment in Education: Czech Republic 2012*. OECD.

Schwarz, C. V., & White, B. Y. (2005). Metamodeling knowledge: Developing students' understanding of scientific modeling. *Cognition and Instruction, 23*(2), 165–205.

Schwarz, C. V., Reiser, B. J., Davis, E. A., Kenyon, L., Acher, A., Fortus, D., et al. (2009). Developing a learning progression for scientific modeling: Making scientific modeling accessible and meaningful for learners. *Journal of Research in Science Teaching, 46*(6), 632–654.

Sedová, K., Svarícek, R., & Salamounová, Z. (2012). Komunikace ve skolní tríde [Communication in the classroom]. *Praha: Portál.*

Slavík, J. (2003). Autonomní a heteronomní pojetí školního hodnocení – aktuální problem pedagogické teorie a praxe [Autonomous and heteronomous assessment – current issue of educational theory and practice]. *Pedagogika*, *53*(1), 5–25.

Sluijsmans, D. M. A. (2002). *Student involvement in assessment, the training of peer-assessment skills.* Interuniversity Centre for Educational Research.

Sluijsmans, D. M., Brand-Gruwel, S., & van Merriënboer, J. J. (2002). Peer assessment training in teacher education: Effects on performance and perceptions. *Assessment & Evaluation in Higher Education, 27*(5), 443–454.

Smit, R. (2009). Die formative Beurteilung und ihr Nutzen für die Entwicklung von Lernkompetenz. In *Eine empirische Studie in der Sekundarstufe 1*. Baltmannsweiler: Schneider Verlag Hohengehren GmbH.

Starý, K. (2006). *Sumativní a formativní hodnocení* [Summative and formative assessment]. Portál RVP, www.rvp.cz

Stefani, L. A. (1994). Peer, self and tutor assessment: Relative reliabilities. *Studies in Higher Education, 19*(1), 69–75.

Strijbos, J. W., Narciss, S., & Dünnebier, K. (2010). Peer feedback content and sender's competence level in academic writing revision tasks: Are they critical for feedback perceptions and efficiency? *Learning and Instruction, 20*(4), 291–303.

Strijbos, J. W., & Sluijsmans, D. (2010). Unravelling peer assessment: Methodological, functional, and conceptual developments. *Learning and Instruction, 20*(4), 265–269.

Tiberghien, A. (2011). Conception et analyse de ressources d'enseignement : le cas des démarches d'investigation. In M. Grangeat (Ed.), *Les démarches d'investigation dans l'enseignement scientifique Pratiques de classe, travail collectif enseignant, acquisitions des élèves* (pp. 185–212). Lyon: INRP.

Topping, K. (1998). Peer assessment between students in colleges and universities. *Review of Educational Research, 68*(3), 249–276.

Topping, K. (2013). Peers as a source of formative and summative assessment. In J. H. Mac Millan (Ed.), *SAGE handbook of research on classroom assessment* (pp. 395–412). London: Sage.

Topping, K. J., Smith, E. F., Swanson, I., & Elliot, A. (2000). Formative peer assessment of academic writing between postgraduate students. *Assessment & Evaluation in Higher Education, 25*(2), 149–169.

van Gennip, N. A., Segers, M. S., & Tillema, H. H. (2009). Peer assessment for learning from a social perspective: The influence of interpersonal variables and structural features. *Educational Research Review, 4*(1), 41–54.

Van Steendam, E., Rijlaarsdam, G., Sercu, L., & Van den Berg, H. (2010). The effect of instruction type and dyadic or individual emulation on the quality of higher-order peer feedback in EFL. *Learning and Instruction, 20*(4), 316–327.

Vögeli-Mantovani, U. (1999). SKBF Trendbericht Nr. 3: Mehr fördern, weniger auslesen. Zur Entwicklung der schulischen Beurteilung in der Schweiz. Aarau: Schweizerische Koordinationsstelle für Bildungsforschung.

Walker, A. (2001). British psychology students' perceptions of group-work and peer assessment. *Psychology Learning & Teaching, 1*(1), 28–36.

Yang, M., Badger, R., & Yu, Z. (2006). A comparative study of peer and teacher feedback in a Chinese EFL writing class. *Journal of Second Language Writing, 15*(3), 179–200.

Žlábkova, I., & Rokos, L. (2013). Pohledy na formativní a sumativní hodnocení žáka v českých publikacích [Formative and summative assessment in Czech publications]. *Pedagogika, 58*(3), 328–354.