Pinger, Petra; Rakoczy, Katrin; Besser, Michael; Klieme, Eckhard

The interplay of formative assessment and instructional quality. Interactive effects on students' mathematics achievement

formal und inhaltlich überarbeitete Version der Originalveröffentlichung in:
formally and content revised edition of the original source in:
Learning environments research 21 (2018) 1, S. 61-79

Bitte verwenden Sie in der Quellenangabe folgende URN oder DOI /
Please use the following URN or DOI for reference:
urn:nbn:de:0111-pedocs-174072
10.25656/01:17407

https://nbn-resolving.org/urn:nbn:de:0111-pedocs-174072
https://doi.org/10.25656/01:17407

This is a post-peer-review, pre-copyedit version of an article published in Learning environments research. The final authenticated version is available online at: http://dx.doi.org/10.1007/s10984-017-9240-2

The Interplay of Formative Assessment and Instructional Quality – Interactive Effects on Students' Mathematics Achievement

Petra Pinger[a *], Katrin Rakoczy[a], Michael Besser[b], Eckhard Klieme[a]

[a] *Department of Educational Quality and Evaluation, German Institute for International Educational Research (DIPF), Frankfurt, Germany*

[b] *Institut for Mathematics and its Didactics at the Leuphana University of Lueneburg, Germany*

* Corresponding author: Petra Pinger, German Institute for International Educational Research, Department of Educational Quality and Evaluation, Schloßstraße 29, D-60486 Frankfurt am Main, Germany, Phone: +49 (0)69 24708-114, pinger@dipf.de

Acknowledgement

This work was supported by the German Research Foundation under Grant KL 1057/10-3, BL 275/16-3 and LE 2619/1-3.

We would like to thank Malte Klimczak for his support in planning and conducting the study.

The Interplay of Formative Assessment and Instructional Quality – Interactive Effects on Students’ Mathematics Achievement

Formative assessment is considered a promising teaching practice to promote teaching and learning processes. The implementation of teaching practices into instruction involves intervening with a learning environment that is characterized by certain features of instructional quality. Our study aims to contribute to the understanding of formative assessment by analysing the interplay between a formative assessment intervention and aspects of general instructional quality. In a quasi-experimental study design, fifteen teachers participated in a control group (n=361 students) and twenty teachers in the intervention classes (n=498 students) implemented a curriculum-embedded formative assessment tool in their ninth grade mathematics classes. No effects were found of the intervention on the assessed aspects of general instructional quality (process-oriented instruction, teacher-student relationship, effective use of instructional time). However, multilevel regression analyses revealed an interaction effect between the intervention and process-orientation and the effective use of instruction time. Our findings suggest that implementing formative assessment tools do not seem to suffice regarding changes in general instructional quality, but that an intervention with detailed material and guidelines can counterbalance effects of instructional quality, fostering students’ achievement in classes with lower degrees of process orientation and a less effective use of instructional time.

Keywords: formative assessment; teaching quality; mathematics; intervention; feedback; instructional quality; achievement; instruction

1. Introduction

The central aim of educational research is to improve teaching and learning processes. One teaching method that claims to achieve this objective is formative assessment (e.g, Black & Wiliam, 2009). Formative assessment is based on the idea of evaluating students’ understanding regularly throughout a teaching unit and making use of this information to improve teaching and learning processes. Teachers can use this assessment information to adapt instruction in accordance to the students' needs and can provide feedback to improve students' learning processes. Formative assessment practices can be implemented successfully in everyday instruction through professional development programs that are supported by predesigned formative assessment material (e.g., Hondrich, Hertel, Adl-Amini, & Klieme, 2016; Schneider & Randel, 2010) but effect sizes have varied greatly across studies indicating that the effectiveness of formative assessment interventions depends on how it is realized (Kingston & Nash, 2011). When a new teaching method is implemented into the classroom, it is brought into a learning environment that is characterized by certain instructional features. In accordance with Fraser (1998) the learning environment of a classroom refers to the "social, psychological and pedagogical contexts in which learning occurs" (p. 3). The classroom as a learning environment can be characterized by structural features of the classroom (e.g. class size, lesson duration) and by aspects of instructional quality. Aspects of the instructional quality like the classroom climate are assumed to impact the effectiveness of formative assessment interventions (McMillan, 2010). At the same time, a new teaching method is implemented to make changes in instruction and thereby to improve the instructional quality. The implementation of teaching practices involves intervening with the learning environment that exists in the classroom, and it is therefore of major interest to evaluate the effects of this interplay between an intervention and existing instructional quality. In this study, we want to contribute to this understanding by analysing the interplay between a formative assessment intervention and aspects of the general instructional quality. In the following, we give a brief outline of theoretical and empirical findings on the effects of general instructional quality and curriculum-embedded formative assessment practices. Then, we discuss the theoretically assumed interplay between formative assessment practices and aspects of the general instructional quality. On this basis, research questions and hypotheses are deduced.

1.1. Instructional quality

When a teaching method is implemented in the existing learning environment of a classroom, it intervenes in a system that is characterized by certain instructional features. One way of describing the learning environment of a classroom is in terms of instructional quality. Drawing on theories of learning and motivation, multiple theoretical frameworks describe instructional quality by three global dimensions: cognitive activation, supportive climate, classroom management (e.g., Fauth, Decristan, Rieser, Klieme, & Büttner, 2014; Klieme, Pauli, & Reusser, 2009; Lipowsky et al., 2009; see also Pianta & Hamre, 2009). The dimension of *cognitive activation* features aspects of the instruction that promote the depth of students' cognitive engagement with the subject matter (e.g., Klieme et al., 2009). Following the ideas of constructivist learning theories (e.g., Cobb, 1994: Dewey, 1916; von Glasersfeld, 1995), instruction is cognitively activating when it "encourages students to engage in (co)-constructive and reflective higher level thinking" (Klieme et al., 2009, p. 140). This includes the use of demanding questions, tasks, and problems that activate students' prior knowledge and stimulate higher-level thinking (e.g., Klieme et al., 2009). The dimension of *supportive climate* is characterized by a warm and caring teacher-student relationship and student-oriented individual support. Drawing on self-determination theory as one of the most prominent motivational theories (e.g., Ryan & Deci, 2000) it is assumed that a supportive learning climate fosters students' motivation and interest (e.g., Fauth et al., 2014; Klieme et al., 2009; Kunter, 2005; Lipowsky et al., 2009; Rakoczy, 2008). The cognitive dimension and the motivational dimension are supplemented by the dimension of *classroom management*. Effective management and prevention of disciplinary problems is relevant for on-task behaviour in class and is thus seen as the prerequisite for high-quality motivational and cognitively activating learning activities (e.g., Klieme et al., 2009; Oliver, Wehby, & Daniel, 2011; Rakoczy et al., 2007; Seidel & Shavelson, 2007; Wang, Haertel, & Walberg, 1993).

The three-dimensional framework enables a differentiated description of generic aspects of instructional quality and provides a theoretical basis for empirical research to predict cognitive and motivational learning outcomes (e.g., Fauth et al., 2014; Kunter, 2005; Lipowsky et al., 2009; Pianta & Hamre, 2009; Rakoczy, 2008).

1.2. Formative assessment

Formative assessment can be understood as a teaching method that aims at improving students' learning by eliciting information about students' understanding and making use of this information to alter teaching and learning processes (e.g., Bloom, Hasting & Madaus, 1971; Black & Wiliam, 1998a; 1998b; 2009; Cizek, 2010). According to research syntheses and meta-analyses, formative assessment seems to fulfil this objective of positively affecting students learning outcomes (Black & William, 1998a; Kingston & Nash, 2011). A great variability of effect sizes found in these studies on formative assessment indicates that effectiveness of a formative assessment intervention depends on how formative assessment is realised and implemented (Kingston & Nash, 2011; William, 2010). In all cases, implementing formative assessment implies cycles of eliciting information through assessments, interpreting the information gathered and acting on behalf of this information to improve teaching and learning processes (e.g., Wiliam, 2010). The assessment itself can be teacher-directed, self-assessment and peer-assessment; it can be oral and written and vary in degrees of formality ranging from spontaneous classroom questioning to more formal assignment (Shavelson et al., 2008; Cizek, 2010). For the assessment to become formative and to complete the formative assessment cycle, it is essential that the assessment information is subsequently used to alter students' learning processes (Black & Wiliam, 1998b; 2009; Sadler, 1989). Providing students with feedback constitutes a powerful tool for altering learning processes and consequently a key strategy in realizing formative assessment (e.g. Hattie & Timperley, 2007; Wiliam & Thompson, 2008). Research on the effectiveness of feedback on learning has shown that not all types of feedback are equally effective (e.g., Kluger & DeNisi, 1996). In the literature, feedback is considered to support cognitive and motivational learning processes when it informs the learner not only about the correctness of a task solution but also about next steps in learning to reach a certain learning goal (e.g. Shute, 2008, Hattie & Timperley, 2007; Sadler, 1989). Formative assessment feedback should be designed in a way that teacher and student receive a clear picture on *where the learner is going*, *where the learner is right now* and *how to get there* (Hattie & Timperley, 2007, Wiliam, 2010). One way of implementing formative assessment is through curriculum-embedded formative assessment intervention (e.g., Shavelson et al., 2008). Curriculum-embedded assessments are assignments designed as part of a teaching unit to check students' understanding at critical junctures within the teaching unit. The advantage of these pre-structured programs is that teachers can be supported by predesigned assessment tools facilitating successful implementation (e.g., Hondrich et al., 2016). The formative feedback the students receive as part of the curriculum-embedded formative assessment can also be predesigned to help teachers provide constructive feedback. Rakoczy and colleagues (2013) and Harks and colleagues (2014) could show under laboratory conditions that a structured feedback tool that includes information on processes that have been mastered (strengths), areas that need further improvement (weaknesses), and recommendations on how to improve (strategies) affects students' learning processes positively via perceived usefulness and perceived competence support.

1.3. Interplay between teaching quality and formative assessment

From a theoretical point of view, the elements of the teaching method 'formative assessment' are closely linked to features of instructional quality. Assessment tasks can be challenging, activate students' prior knowledge and engage students in constructive learning processes and therefore make a contribution to cognitively activating instruction (e.g., Baumert et al., 2010;

Klieme et al., 2009). Furthermore, feedback is not only an essential element of formative assessment but is also an aspect of individual learners' support and part of a supportive climate (e.g. Klieme et al., 2009, Lipowsky et al., 2009). The overlap of formative assessment practices with indicators of global dimensions of instructional quality is also apparent in studies on the quality of enactment of formative assessment practices. Furtak and colleagues (2008) for example used the following criteria to evaluate enactment of formative assessment: Eliciting student conceptions, tracking and clustering student conceptions, asking students to provide reasons for their explanations, students argue ideas and evidence, and students provide evidence for their claims. Similar criteria could be used as indicators for the concept of cognitive activation (e.g. Lipowsky et al., 2009). Consequently, two conceivable links can be assumed concerning the interplay between a formative assessment intervention and instructional quality.

First, formative assessment aims at improving not only learning processes but also teaching processes (e.g. Black & William 1998b). Insofar high levels of teaching quality can be seen as a positive outcome of formative assessment. The information provided by the assessment is supposed to equip teachers with knowledge about the level of their students' understanding (e.g., Wiliam, 2010). This may enable the teachers to assign appropriate tasks with a degree of difficulty that is challenging and matches the students' prior knowledge, leading to a cognitively activating learning environment. Moreover, constructive feedback as a central element of formative assessment may lead to a supportive learning climate. Giving students formative feedback that causes an increased sense of competence (Rakoczy, Harks, Klieme, Blum, & Hochweber, 2013) might strengthen the relationship among teacher and students, and contribute to the development of an environment that students perceive to be supportive. Moreover, the predesigned assessment and feedback material and the associated clear structure of the intervention may facilitate an effective management of time.

Consequently, implementing formative assessment may help teachers to improve aspects of their instructional quality.

Second, two types of interactive effects of formative assessment and the instructional quality on students learning outcomes are plausible. A positive interaction would indicate that aspects of instructional quality are seen as reinforcement of the effects of formative assessment practices (e.g., Decristan et al., 2015). A high degree of classroom management might be a relevant precondition for a high degree of on-task behaviour and the engagement with assessment tasks and feedback information. In classroom environments that are characterized as cognitively activating, students and teachers routinely discuss solution processes at a high level and debate multiple task solutions (e.g., Klieme et al., 2009). This might facilitate engagement with formative assessment tasks and feedback. A warm and caring teacher-student relationship and a perceived positive error culture might increase the chance that students make use of the formative feedback provided by their teachers. A supportive classroom climate in which students are comfortable giving and receiving feedback and discussing their ideas might be crucial for successful formative assessment (McMillan, 2010; Sadler, 1989). This theoretical link was recently studied by Decristan and colleagues (2015) who conducted an empirical study, and found positive effects on students' achievement of both their formative assessment intervention and aspects of the instructional quality. Moreover, they found a positive interaction between formative assessment and the instructional quality dimensions of cognitive activation and supportive climate. Hence, they concluded that for instruction to become most effective, specific teaching practices must be combined with high instructional quality.

On the other hand, it is supposable and desirable that a specific teaching method that is implemented in classrooms supports teachers with low degrees of general instructional quality. In the case of formative assessment, a predesigned feedback tool that combines characteristics of good feedback known from the research literature might counterbalance the effects of a less supportive climate. Cognitively demanding assessment tasks and process-oriented feedback might compensate for lessons that are usually less cognitively activating. Moreover, the generally high degree of prestructuredness of a formative assessment intervention might compensate for a less effective use of instructional time.

2. Research questions

Drawing on the theoretical and empirical findings in the literature on formative assessment and feedback discussed in the previous sections, two research questions can be formulated.

1. Effects of formative assessment on instructional quality: Does an intervention based on curriculum-embedded formative assessment practices affect the general instructional quality (cognitive activation, supportive climate, classroom management)?
2. Interaction effects of curriculum embedded formative assessment and instructional quality: Is the effect of the general instructional quality on achievement moderated by the formative assessment intervention?

3. Method

The intervention study was realized in the academic year 2010/2011 as part of the project 'Conditions and Consequences of Classroom Assessment (Co²CA)' which was conducted by the German Institute for International Educational Research, the University of Kassel and the University of Lüneburg, and was funded by the German Research Foundation[1]. The description of the project and of the study design is limited to the part relevant to analyses presented here. The design of the complete project is reported in Rakoczy, Klieme, Leiss and Blum (2017). The study presented here is part of a larger quasi-experimental study

[1] The project was supported by grants from the German Research Foundation (DFG, KL 1057/10-3, BL 275/16-3 and LE 2619/1-3); principal researchers: E. Klieme, K. Rakoczy (both Frankfurt), W. Blum (Kassel), D. Leiss (Lüneburg).

investigating the impact of formative assessment interventions on learning. The intervention groups implementing written curriculum-embedded formative assessments were combined in the present analyses.

3.1. Participants

Thirty-nine teachers from middle track state schools in Hesse, Germany (urban and rural areas) participated in this intervention study with their ninth grade mathematics classes. Participation in the study was voluntary for teachers and students and written informed consent of the parents was obtained for all participating students.

We excluded four classes from the intervention groups due to drop-outs caused by illness or incomplete adherence to intervention guidelines. The final sub-sample consisted of 35 teachers and 859 students. For the cluster randomized field trial with pre-test and post-test, classes were randomly assigned to the control group (n=15) or the intervention groups (n=20). Teachers in the control group originated from 11 schools; they ad on average 22 years of teaching experience, and 66.7% were female. Teachers in the intervention groups originated from 13 schools; they had on average 17 years of teaching experience, and 68.4% were female. The students in both conditions had a mean age of 15.1 years, and (46.6 % were female).

3.2. Design and Procedure

The treatment conditions were realized through teacher trainings. In order to keep the teaching content constant across the participating classes, all teachers in the control and intervention groups were asked to teach the same predesigned teaching unit. The teaching unit consisted of 13 lessons (45 minutes each) and had four phases: (1) an introduction including a proof and technical tasks, (2) word problems, (3) modelling problems, and (4) consolidation. To keep instruction as consistent as possible, all teachers received detailed guidelines which included a description of the teaching unit and a description of learning goals to be achieved in each phase. Additionally, teachers were given illustrations of obligatory teaching material to assure that all students worked on the same tasks. In this first teacher training session, all teachers received general organisational information concerning the study and subject-specific content concerning the predesigned teaching unit on Pythagoras' theorem. The teachers in the intervention groups received a second teacher training on formative assessment practices and were trained to assess their students' understanding at three critical junctures within the teaching unit. At the end of phases one to three (at the end of lessons 5, 8, and 11), teachers were asked to administer a diagnostic tool to assess students' understanding and to provide feedback. The diagnostic tools consisted of two components: (1) assessment: one or two mathematical problems and space for the student to write down the solution, (2) process-oriented feedback: three text-fields to indicate strengths, weaknesses, and strategies to improve (see Figure 1 for an example of the second diagnostic tool). As the name implies, process-oriented feedback focuses on the processes and operations needed to complete tasks. The feedback includes information on processes that have been mastered (strengths), areas that need further improvement (weaknesses), and recommendations on how to improve (strategies; see Harks, Rakoczy, Hattie, Besser, & Klieme, 2014 and Rakoczy et al., 2013 for more information on the effects of process-oriented feedback and Pinger, Rakoczy, Besser, & Klieme, 2016 for more details on the implementation in this study). Teachers returned the corrected diagnostic tool with process-oriented feedback in the subsequent lesson (beginning of lessons 6, 9, and 12).

Prior knowledge was assessed immediately before the teaching unit and achievement in mathematics was assessed immediately afterwards. Instructional quality was assessed via student questionnaires before and after the teaching unit (see Figure 2 for an overview of the study design).

3.3. Measures

3.3.1. Achievement

Before administering the predesigned teaching unit, students' prior knowledge was assessed using a pre-test (19 items); in the lesson immediately after the teaching unit mathematical achievement on Pythagoras' theorem was assessed by a post-test (17 items). The pre-test did not assess knowledge of Pythagoras' theorem but focused on relevant prior knowledge such as identifying a right-angled triangle and solving equations. The post-test included technical tasks, word problems, and modelling tasks. The items had been analysed previously in a scaling study (N = 1570; Harks, Klieme, Hartig, & Leiss, 2014); therefore, item parameters from the one-dimensional Rasch model of the scaling study could be used as fixed parameters. Weighted likelihood estimator parameters served as achievement scores for pre-tests and post-tests. Estimated reliability (EAP/PV) was .66 for the pre-test and .74 for the post-test. Additionally we computed intraclass correlations (ICC1s) as indicators of the proportion of total variance that can be attributed to between-class differences. ICC1s for pre-test and post-test were .17 and .10, respectively.

3.3.2 Instructional Quality

Cognitive activation – process-oriented instruction

Instruction becomes cognitively activating not only by administering challenging tasks but most importantly through dealing with tasks in a challenging way. Accordingly, we used a scale capturing the process-oriented dealing with tasks and problems in instruction (Klimczak et al, 2012; adapted from Rakoczy, Buff, & Lipowsky, 2005). The scale consists of five items and students were asked to evaluate on a four-point scale ranging from 1 (completely disagree) to 4 (completely agree) if the teacher was interested in and focused on the process of solving tasks and problems ("My mathematics teacher is interested in the way we solve problems."; "My mathematics teacher likes it when we find new approaches to solve a problem”). Internal consistency of the scale was Cronbach’s α .78 for the pre-questionnaire and .84 for the post-questionnaire. As a reliability measure of the class average, we computed ICC2s (e.g., Lüdtke, Robitzsch, Trautwein, & Kunter, 2009). Based on an average class size of 22 students, ICC2s for pre-questionnaire and post-questionnaire were .86 and .79, respectively. ICC1s were .22 and .15, respectively.

Supportive climate – teacher-student relationship

For the dimension of supportive climate, we assessed the relationship between teacher and students (Klimczak et al, 2012; Rakoczy, Buff, & Lipowsky, 2005; originally adapted from TIMSS+, 1995; Prenzel, Kirsten, Dengler, Ettle, & Beer, 1996). On a four- point scale ranging from 1 (completely disagree) to 4 (completely agree), students rated 7 items related to the teacher-student relationship (i.e., “I have the feeling that the teacher takes me seriously”; “I think my teacher likes me”). Cronbach’s α was .84 for the pre-questionnaire and .81 for the postquestionnaire, respectively. ICC2s for pre- and post-questionnaires were .81 and .80. ICC1s were .17 and .16, respectively.

Classroom management – effective use of instructional time

The classroom management scale referred to the maximal use of instructional time (Klimczak et al, 2012; Rakoczy, Buff, & Lipowsky, 2005; Waldis, Buff, Pauli, & Reusser, 2002). Students rated four items ranging from completely disagree (1) to completely agree (4) relating to the time it takes until the lesson can begin and the amount of class time that is idled away (“It usually takes a long time until everybody is quiet and we can start with the lesson”). All items were recoded before scaling. Cronbach’s α were .74 and .73 and ICC2s were .90 and .84 for pre- and post-questionnaires, respectively. ICC1s were .30 and .30, respectively.

3.4. Data analysis

Before conducting simple and multilevel regression analyses in Mplus7 (Muthén & Muthén, 2012), we prepared the data in SPSS 20. Student ratings of the measures of the three aspects of general instructional quality were aggregated. We z-standardized scores of prior knowledge, post-test achievement and instructional quality at the student level (level 1). The aggregated instructional quality scores were standardized at the classroom level (level 2).

Our first research question refers to the effects of the curriculum-embedded formative assessment intervention on the aspects of the general instructional quality, analysed by linear regressions at the classroom level. In three separate analyses, we used the aggregated student ratings of the post-questionnaire as outcome variables, the aggregated pre-questionnaire scores as covariates and the dummy-coded treatment variable as predictor.

For all further analyses regarding research question 2, we conducted multilevel regression analyses to account for the nested data structure. In all cases, students' post-test achievement scores formed the outcome variable and we included pre-test score as covariate at the student level. Before analysing the interactive effects, we tested for main effects of the intervention and the aspects of the general instructional quality. The effect of our intervention on student achievement was tested by including the dummy coded treatment variable on level 2 (0 = control group, 1 = formative assessment). Concerning the main effects of teaching quality, we ran separate regression analyses for each aspect of instructional quality (process orientation, teacher-student relationship, use of lesson time). Aggregated and z-standardized ratings of instructional quality were included as independent variables on the classroom level. Additionally, we included the group-mean centred individual ratings (e.g., Lüdtke et al., 2009). The interaction effects were addressed by extending the models by including the intervention dummy variable and a classroom-level interaction variable that is the product of the dummy coded intervention variable and the aggregated student ratings of instructional quality. We analysed interactive effects on the basis of the pre-questionnaire measures since we were interested in the effects of instructional quality as it originally existed and the interactive effects of the intervention implemented in the existing learning environment.

4. Results

Table 1 presents the descriptive data including correlations sample sizes, means and standard deviations. Correlations between individual-level variables are listed below the diagonal and correlations between class-level aggregated variables are listed above the diagonal. In general, correlations between variables at the classroom level were greater than the correlations at the individual level. Besides the high correlations between pre-test and post-test measures, high correlations were found between the measures of instructional quality, especially between process-orientation and teacher-student relationship. The mean score of use of instructional time at measurement point 2 is higher ($M = 2.67$) compared to the first measurement point ($M = 2.33$). The mean scores for process-orientation and for teacher-student relationship are slightly lower at the second measurement point ($M = 3.28$ and $M = 3.13$, $M = 2.93$ and $M = 2.80$, respectively). Dependent sample t-tests conducted on the class level revealed that all three differences are statistically significant (use of time: $t(34) = -8.31$, $p < .01$; process-orientation: $t(34) = 6.96$, $p < .01$, teacher-student relationship: $t(34) = 5.69$, $p < .01$).

Table 1. Descriptives.

Variable		1	2	3	4	5	6	7	8	9	*n*	*M*	*SD*
1	achievement T1	-	.49**	-.08	-.11	-.14	-.21	.01	.02	.02	762	-1.11	.91
2	achievement T2	.45**	-	.33	.19	.26	.14	.29	.21	.10	783	-.17	1.14
3	process-orientation T1	.02	.02	-	.91**	.86**	.88**	50**	.53**	.14	757	3.28	.57
4	process-orientation T2	.02	.10**	.43**	-	.76**	.87**	.42*	.52**	.19	779	3.13	.63
5	teacher-student relation T1	.04	.06	.67**	.40**	-	.91**	.45**	.43*	.06	758	2.93	.63
6	teacher-student relation T2	-.01	.06	.45**	.71**	.55**	-	.35*	.42*	.17	778	2.80	.61
7	use of time T1	-.03	.04	.16**	.11**	.22**	.14**	-	.81**	-.15	755	2.33	.71
8	use of time T2	-.03	.02	.17**	.21**	.19**	.28**	.47**	-	-.22	780	2.67	.67
9	intervention [a]	-.05	.02	.09*	.09*	.05	.09*	-.07	-.09**	-	859		

Note. Correlations between individual-level variables are listed below the diagonal; correlations between class-level aggregated variables are listed above the diagonal. Means and standard deviations refer to the individual-level variables.

*p < .05. **p <.01; [a] dummy-coded: control = 0 and intervention = 1

Regarding the first research question, we tested the effect of the formative assessment intervention on instructional quality. For all three aspects of instructional quality, no statistically significant effect could be found (Table 2). The regression analysis regarding the main effects of our intervention indicates that control group and intervention group did not differ in post-test achievement when pre-test scores are controlled for (Table 3; Model 4).

Table 2. Regression analysis predicting teaching quality from the intervention

	Model 1			Model 2			Model 3		
	β	*SE*	*p*	*β*	*SE*	*p*	*β*	*SE*	*p*
Process-orientation T1	.899	.071	.000						
Teacher-student relationship T1				.898	.069	.000			
Use of time T1							.796	.098	.000
Intervention [a]	.122	.141	.387	.226	.138	.103	-.202	.196	.301

Note. [a] Dummy-coded: control = 0 and intervention = 1.

Table 3. Multilevel regression analyses predicting student achievement from intervention and teaching quality.

	Model 4			Model 5			Model 6			Model 7		
	β	*SE*	*p*	*β*	*SE*	*p*	*β*	*SE*	*p*	*β*	*SE*	*p*
Individual level												
Pre-test	.445	.043	.000	.463	.042	.000	.455	.043	.000	.459	.042	.000
Process-orientation				-.061	.038	.108						
Teacher-student relationship							.002	.039	.950			
Use of instructional time										-.006	.046	.893
Class-level												
Intervention [a]	.119	.123	.333									
Process-orientation				.130	.071	.067						
Teacher-student relationship							.110	.063	.082			
Use of instructional time										-.086	.069	.212

Note. [a] Dummy-coded: control = 0 and intervention = 1.

Table 4. Multilevel regression analyses predicting student achievement from intervention, teaching quality, and their interactions.

	Model 8			Model 9			Model 10		
	β	*SE*	*p*	*β*	*SE*	*p*	*β*	*SE*	*p*
Individual level									
Pre-test	.472	.042	.000	.460	.043	.000	.463	.042	.000
Process-orientation	-.061	.038	.107						
Teacher-student relationship				.002	.039	.956			
Use of instructional time							-.007	.046	.887
Class-level									
Intervention	.067	.112	.550	.105	.115	.362	.162	.106	.128
Process-orientation	.277	.098	.005						
Teacher-student relationship				.197	.120	.102			
Use of instructional time							.259	.115	.024
Interactions									
Intervention * Process-orientation	-.247	.106	.020						
Intervention * Teacher-student relationship				-.141	.128	.271			
Intervention * use of instructional time							-.262	.125	.036

Note. [a] Dummy-coded: control = 0 and intervention = 1.

We did not find any statistically significant effects of all three aspects of instructional quality measured prior to our intervention on students' post-test achievement (Table 3; Model 5-7). In regard to the second research question, we tested for interaction effects between the formative assessment intervention and instructional quality by including an interaction term in the multilevel regression analyses. As shown in Table 4 (Model 8-10), we found a negative interaction for the aspect of process orientation and for use of instructional time. The statistically significant positive coefficients of process-orientation and use of instructional time in Model 8 and 10 indicate that there is a positive association between these two aspects of instructional quality and students' achievement. However, the negative coefficients of the interactions in the same models indicate that this positive association is suppressed by the formative assessment intervention.

5. *Discussion*

Our study aims to contribute to the understanding of the interplay between the implementation of curriculum-embedded formative assessment practices and general instructional quality as it exists prior to the intervention. First, we looked into the question of how implementing curriculum-embedded formative assessment affects aspects of the general instructional quality. We hypothesised that a curriculum-embedded formative assessment intervention that guides teachers to use predesigned assessment tasks and process-oriented feedback would improve the perceived process orientation of instruction, the teacher-student relationship and the perception of an effective use of instructional time. Contrary to our expectations, no statistically significant effects could be found. Our teacher training enabled teachers to implement successfully the diagnostic tools that consisted of challenging tasks and feedback focusing on strengths, weaknesses and hints in the solution process (see also Pinger et al., 2016). However, our results indicate that implementing challenging tasks and supportive feedback via predesigned tools does not automatically change the general instructional quality. In line with this finding, a subsequent study on the effects of professional development programs that extended the formative assessment intervention described in this study revealed that a professional development program can foster the teachers' pedagogical content knowledge without increasing the general teaching quality (Besser et al, 2016, in preparation). Improving instructional quality includes making profound changes in professional routines and because making these changes is challenging, it seems plausible that our intervention and especially the teacher training was too short-term to initiate improvement (e.g., Creemers, Kyriakides, & Antoniou, 2013; Desimone, 2009). In designing the teacher trainings we were faced with the problem of finding the right trade-off between additional time and work load for the participating teachers and the intensiveness of the training. Both teacher trainings lasted about five hours. The teacher training took into account two important aspects of professional development. First, the trainings were supported by video material. Teachers watched and discussed short video clips of real classroom situations. The use of videos to initiate reflection and discussion upon teaching practices is seen as a valuable element of professional development (e.g., Blomberg, Sherin, Renkl, Glogger, & Seidel, 2014; Borko, Jacobs, Eiteljorg, & Pittman, 2008). Second, our teacher trainings were complemented by detailed guidelines and material. A clear content focus and supportive material is another important aspect of effective professional development (e.g., Desimone, 2009; Hondrich et al., 2016). However, the teacher training might not have included enough active learning possibilities and was not spread across a sufficiently long time period to affect changes in general instructional routines (Desimone, 2009; Schneider & Randel, 2010). A linear relationship between training duration and its effect is not assumed, but short term trainings of less than 20 hours are supposed to be less effective than longer-term professional development programs that are spread over a longer period of time (e.g., Desimone, 2009).

Surprisingly, we found changes in the instructional quality from the pre-questionnaire to the post-questionnaire for the complete sample of all 35 classes (control and intervention groups). The use of instructional time increased while process orientation and teacher-student relationship decreased. The increase in use of instructional time is probably caused by detailed guidelines that all teachers received regardless of the condition. Detailed information about learning phases, corresponding learning goals, and supplemental material provided structural support and consequently might have increased the use of instructional time. Besides, in the intervention period students did not receive homework and administering and controlling homework usually also takes up instructional time. The decrease in teaching quality might indicate that "something new" is not automatically perceived as "something good". First of all, "something new" is "something different". Generally, people show the tendency to prefer things staying as they are, known as the status quo bias in research on decision making (Kahneman, Knetsch, & Thaler, 1991). The predesigned teaching unit that all teachers were asked to teach was designed to foster competency-oriented teaching by focusing not only on technical competencies but also on modelling competencies (Besser, M., Blum, W. & Klimczak, M., 2013). By means of modelling problems, students have to apply their newly developed knowledge and skills to real-world problems (e.g., Niss, 2003; Bloomhoj & Jensen, 2007). The students participating in our study might not be used to competency–oriented instruction (Kunter et al., 2013) and more critical towards a new way of teaching that apparently deviates from their normal classroom routines.

As a second possible type of interplay between formative assessment and instructional quality, we were interested in possible interactive effects. We supposed that either a high degree of instructional quality could be a reinforcement of the effects of our intervention resulting in a positive interaction, or that curriculum-embedded formative assessment as realized in our intervention counterbalances effects of the general instructional quality. As a first step, we looked into the main effects of our intervention and the main effects of the assessed aspects of instructional quality (process orientation, teacher-student relationship, use of instructional time). The intervention and the three aspects of instructional quality did not show statistically significant effects on students' mathematics achievement. Concerning the interactive effects, no interaction effect was found between our intervention and the teacher-student relationship that was used as an indicator of the supportive learning climate. However, we found a negative interaction between our formative assessment intervention and process orientation and the use of instructional time. The negative interaction indicates that the assumed positive association between process-orientation and use of instructional time and achievement is supressed by the implementation of the formative assessment. The increased structure of repeated formative assessment at critical junctures within the teaching unit and formative feedback focusing on providing hints for improvement seem to compensate for low degrees of efficient use of instructional time and process orientation. The formative assessment tool probably helps teachers who lose instructional time due to low degrees of classroom management to make use of the available time most effectively. The process-oriented feedback tools, with predesigned fields for strengths, weaknesses and hints, might counterbalance for less process-oriented instruction, i.e. lessons in which teachers focus less on the way students solve problems and in which students' ideas of new approaches to solving a problem are less frequently discussed. At the same time, the negative interaction indicates that students profit less from implementing a formative assessment intervention in instruction that is already characterized by high degrees of process orientation and efficient use of instructional time. Our intervention was supplemented by detailed guidelines and predesigned material. This structure helped teachers in instruction with a low degree of instructional quality but might place too high constraints on teachers whose instruction already reveals high instructional quality. The finding of negative interactive effects found in our study contradicts findings from a previous study in which a positive interaction was found (Decristan et al., 2015). In their study, the authors implemented a formative assessment intervention in primary school science classes and found a positive interaction between formative assessment and teaching quality. Decristan and colleagues (2015) used post-questionnaire student ratings of instructional quality and used the intervention period as reference for the student ratings. Hence, their study focused on the interactive effects within the intervention period while we were interested in interactive effects between formative assessment and quality features of instruction as given prior to the intervention. As formative assessment practices are closely related to instructional quality, it cannot be ruled out that students' perception of the instructional quality measured after the intervention was biased by the implementation of the formative assessment intervention. To disentangle both effects, we used the students' perception of instructional quality prior to the intervention.

5.1. Limitations and implications for further research

Comparability across groups was increased by asking all teachers in both conditions to teach a predesigned teaching unit. The design of the teaching unit was guided by clear learning objectives – a key element of formative assessment (Wiliam & Thompson, 2008). We do not know how explicitly teachers communicated the learning objectives, but they were taken into account in the design of the teaching unit. Therefore, the control condition also featured important aspects of formative assessment and it might have thus been more difficult to find significant differences between conditions.

Regarding limitations of our findings, it is also necessary to note that our teacher training was short in time. While all necessary information to implement the predesigned teaching unit, assessment tasks and feedback was provided, the professional development session might not have been long and intensive enough to foster changes in general instructional quality.

Moreover, analyses regarding the effects of our intervention on teaching quality were conducted with the aggregated scores on the class level. Due to exclusion (e.g., drop out caused by illness) the sample was limited to fifteen and twenty classes in control group and experimental condition respectively. The generally small sample size might not have sufficed to reveal statistically significant effects.

The pre-test in our study had with .66 only a moderate reliability. The pre-test covered relevant prior knowledge with regard to the teaching unit on the Pythagorean Theorem. The test included items addressing prerequisites for developing technical competencies (extracting a root, solving equations) and prior knowledge relevant for improving modelling competencies (finding a right-angled triangle in a real world situation shown on a picture). Reliability of the test might have been reduced by covering two cognitive domains in one test (Harks, Klieme, Hartig, & Leiss, 2014).

We used student questionnaires to measure aspects of instructional quality. Student perception of the learning environment has been shown to be a relevant predictor of cognitive and affective learning outcomes (e.g., Fraser 2012; Taylor & Fraser, 2013). The aggregated student ratings are generally seen as valid measures of instructional processes (e.g., De Jong & Westerhof, 2001; Fauth et al., 2014). However, they are not free of potential biases. Student ratings, teacher ratings and ratings of external observers provide an overlapping but in some parts unique perspective on the learning environment (Clausen, 2002, Kunter & Baumert, 2006; Maulana & Helms-Lorenz, 2016). In intervention studies, students' pre-post ratings might be biased by a tendency to prefer the old way of teaching to the newly implemented teaching method. In research on decision-making, this phenomenon is known as the status quo effect (Kahneman et al., 1991). Therefore, it would be interesting to repeat the study and to use observer ratings instead of student ratings.

The quantitative analyses presented in this study looked at the interactive effects on an overall level. Further analyses of individual teachers' results including additional teacher, class and student variables could provide a deeper insight into the question of who benefits most from predesigned formative assessment material.

Moreover, further research could also address more or other aspects of general teaching quality. In this study, we focused on process orientation, teacher-student relationship and use of instructional time. A climate that is characterized by a mastery goal orientation and positive error culture are further examples of aspects of the learning environment that are supposed to be relevant with regard to the effectiveness of formative assessment intervention (e.g., McMillan, 2010).

5.2. Concluding remarks

In this study, we investigated the interplay between a formative assessment intervention and characteristics of instructional quality. Our findings contribute to our understanding of formative assessment as they indicate that formative assessment intervention can foster students' achievement in classes that are characterized as showing lower degrees of process orientation and use of instructional time. In these classes, a structured assessment and feedback tool administered regularly throughout a teaching unit supports students' learning. However, our results also indicate that students profit less from a formative assessment intervention in instruction that is already characterized by high degrees of instructional quality.

Our findings suggest that implementing formative assessment tools by means of short-term teacher trainings does not seem to suffice regarding changes in general instructional quality, but that an intervention with detailed material and guidelines can counterbalance effects of instructional quality. In order to find out in which learning environments formative assessment is most effective and how formative assessment can improve instructional quality sustainably, further research is needed to increase our understanding of how to design and implement formative assessment interventions in various learning environments.

6. References

Baumert, J., Kunter, M., Blum, W., Brunner, M., Voss, T. Jordan, A., ... Tsai, J. (2010). Teachers' mathematical knowledge, cognitive activation in the classroom and student progress. *American Educational Research Journal,* 47(1), 133-180.

Besser, M., Blum, W., & Klimczak, M. (2013). Formative assessment in every-day teaching of mathematical modelling: implementation of written and oral feedback to competency- oriented tasks. In G. A. Stillman, G. Kaiser, W. Blum & J. P. Brown (Eds.), *Teaching mathematical modelling: connecting to research and practice* (pp. 469-478). Springer: New York.

Besser, M., Leiss, D., Klieme, E., Blum, W., Rakoczy, K., & Schütze, B. (2016). *The influence of mathematics teacher professional development on pedagogical content knowledge, on beliefs about teaching and learning, and on the quality of teaching.* Manuscript in preparation.

Black, P., & Wiliam, D. (1998a). Inside the black box. *Phi Delta Kappan, 80*(2), 139- 148.

Black, P., & Wiliam, D. (1998b). Assessment and classroom learning. *Assessment in Education, 5*(1), 7-73.

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability, 21,* 5-31.

Blomberg, G., Sherin, M. G., Renkl, A., Glogger, I., & Seidel, T. (2014). Understanding video as a tool for teacher education: Investigating instructional strategies to promote reflection. *Instructional Science*, *42*(3), 443-463. doi:10.1007/s11251-013-9281-6

Bloom, B. S., Hasting, J. T., & Madaus, G. F. (1971). *Handbook on Formative and Summative Evaluation of Student Learning*. New York: McGraw-Hill.

Bloomhoj, M., & Jensen, T. H. (2007). What's all the fuss about competencies? Experiences with using a competence perspective on mathematics education to develop the teaching of mathematical modelling. In W. Blum, P. L. Galbraith, H.-W. Henn, &

M. Niss (Eds.), *Modelling and Applications in Mathematics Education. The 14th ICMI Study* (Vol. 10, pp. 45–56). New York: Springer.

Borko, H., Jacobs, J., Eiteljorg, E., & Pittman, M. E. (2008). Video as a tool for fostering productive discussions in mathematics professional development. *Teaching and Teacher Education*, *24*, 417–436. doi:10.1016/j.tate.2006.11.012.

Cizek, G. J. (2010). An introduction to formative assessment: History, characteristics, and challenges. In H. L. Andrade & G. J. Cizek (Eds.), *Handbook of Formative Assessment* (pp. 3-17). New York, NY: Routledge.

Clausen, M. (2002). *Unterrichtsqualität: Eine Frage der Perspektive?* [Teaching quality: A matter of perspective?]. Münster, Berlin: Waxmann.

Cobb, P. (1994). Where is the mind? Constructivist and sociocultural perspectives on mathematical development. *Educational Researcher*, *23*, 13-20.

Creemers, B.P.M., Kyriakides, L., & Antoniou, P. (2013). *Teacher Professional Development for Improving Quality of Teaching*. Dodrecht: Springer.

Decristan, J., Klieme, E., Kunter, M., Hochweber, J., Büttner, G., Fauth, B., ... Hardy, I. (2015). Embedded formative assessment and classroom process quality: How do they interact in promoting science understanding?. *American Educational Research Journal, 52*(6), 1133-1159.

De Jong, R., & Westerhof, K. J. (2001). The quality of student ratings of teacher behaviour. *Learning Environments Research*, *4*(1), 51–85.

Desimone, L. M. (2009). Improving impact studies of teachers' professional development: Toward better conceptualizations and measures. *Educational Researcher*, *38*(3), 181-199. doi:10.3102/0013189X08331140

Dewey, J. (1916). *Democracy and Education: An Introduction to the Philosophy of Education*. New York: Macmillan

Fauth, B., Decristan, J., Rieser, S., Klieme, E., & Büttner, G. (2014). Student ratings of teaching quality in primary school: Dimensions and prediction of student outcomes. *Learning and Instruction*, *29*, 1–9.

Furtak, E. M., Ruiz-Primo, M. A., Shemwell, J. T., Ayala, C. C., Brandon, P. R., Shavelson, R. J., & Yin, Y. (2008). On the fidelity of implementing embedded formative assessments and its relation to student learning. *Applied Measurement in Education, 21,* 360-389. doi:10.1080/08957340802347852

Fraser B. J. (1998). The birth of a new journal: editor's introduction. *Learning Environments Research*, *1*, 1–5

Fraser, B. J. (2012). Classroom learning environments: Retrospect, context and prospect. In B. J. Fraser, K. G. Tobin, & C. J. McRobbie (Eds.), *Second international handbook of science education* (pp. 1191–1239). New York: Springer.

Harks, B., Klieme, E., Hartig, J., & Leiss, D. (2014). Separating cognitive and content domains in mathematical competence. *Educational Assessment*, *19*(4), 243-266.

Harks, B., Rakoczy, K., Hattie, J., Besser, M., & Klieme, E. (2014). The effects of feedback on achievement, interest and self-evaluation: the role of feedback's perceived usefulness. *Educational Psychology*, *34*(3), 269-290. doi:10.1080/01443410.2013.785384

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

Hondrich, A. L., Hertel, S., Adl-Amini, K., & Klieme, E. (2016). Implementing curriculum-embedded formative assessment in primary school science classrooms. *Assessment in Education: Principles, Policy & Practice, 23*(3), 353-376. doi:10.1080/0969594X.2015.1049113

Kahneman, D., Knetsch, J. L., & Thaler, R. H. (1991). Anomalies: The endowment effect, loss aversion, and Status Quo Bias. *Journal of Economic Perspectives*, *5*(1), 193–206.

Kingston, N., & Nash, B. (2011). Formative assessment: A meta-analysis and a call for research. *Educational Measurement: Issues and Practice, 30*(4), 28-37.

Klieme, E., Pauli, C., & Reusser, K. (2009). The Pythagoras study: investigating effects of teaching and learning in Swiss and German mathematics classrooms. In T. Janik & T. Seidel (Eds.), *The power of video studies in investigating teaching and learning in the classroom* (pp. 137–160). Münster: Waxmann.

Klimczak, M., Kampa, M., Bü germeister, A., Harks, B., Rakoczy, K., Besser, … Leiss, D. (2012). *Dokumentation der Befragungsinstrumente der Interventionsstudie im Projekt "Conditions and Consequences of Classroom Assessment" (Co²CA)* [Documentation of the survey instruments of the intervention study of the project "Conditions and Consequences of Classroom Assessment" (Co²CA)]. Frankfurt am Main: DIPF 2012.

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119*, 254-284. doi:10.1037//0033-2909.119.2.254

Kunter, M. (2005*). Multiple Ziele im Mathematikunterricht [Multiple goals in Mathematics instruction]*. Münster: Waxmann.

Kunter, M., & Baumert, J. (2006). Who is the expert? Construct and criteria validity of student and teacher ratings of instruction. *Learning Environments Research*, *9*, 231-251. doi:10.1007/s10984-006-9015-7

Kunter, M., Baumert, J., Blum, W., Klusmann, U., Krauss, S., & Neubrand, M. (2013). *Cognitive activation in the mathematics classroom and professional competence of teachers. Results from the COACTIV project*. New York, NY: Springer.

Lipowsky, F., Rakoczy, K., Pauli, C., Drollinger-Vetter, B., Klieme, E., & Reusser, K. (2009). Quality of geometry instruction and its short-term impact on students' understanding of the Pythagorean Theorem. *Learning and Instruction, 19,* 527-537.

Lüdtke, O., Robitzsch, A., Trautwein, U., & Kunter, M. (2009). Assessing the impact of learning environments: How to use student ratings of classroom or school characteristics in multilevel modeling. *Contemporary Educational Psychology*, *34*(2), 120-131.

Maulana, R., & Helms-Lorenz, M., (2016). Observations and student perceptions of the quality of preservice teachers' teaching behaviour: construct representation and predictive quality. *Learning Environments Research*, 19, 335–357. doi:10.1007/s10984-016-9215-8

McMillan, J. H., (2010). The practical implications of educational aims and contexts for formative assessment. In H. L. Andrade & G. J. Cizek (Eds.), *Handbook of Formative Assessment* (pp. 41-58). New York, NY: Routledge.

Muthén, L. K., & Muthén, B. O. (1998–2012). Mplus (Version 7) [Computer Software]. Los Angeles, CA: Muthén & Muthén.

Niss, M. (2003). Mathematical competencies and the learning of mathematics: the danish KOM project. In A. Gagatsis & S. Papastavridis (Eds.), *Mediterranean Conference on Mathematical Education* (pp. 115–124). Athen: 3rd Hellenic Mathematical Society and Cyprus Mathematical Society.

Oliver, R. M., Wehby, J. H., & Reschly, D. J. (2011). Teacher classroom management practices: effects on disruptive or aggressive student behavior. *Campbell Systematic Reviews* 4, doi: 10.4073/csr.2011.4

Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Reseacher 38*(2), 109-199. doi:10.3102/0013189X09332374

Pinger, P., Rakoczy, K., Besser, M., & Klieme, E. (2016). Implementation of formative assessment – effects of quality of programme delivery on students' mathematics achievement and interest. *Assessment in Education: Principles, Policy & Practice*. Advance online publication. doi:10.1080/0969594X.2016.1170665.

Prenzel, M., Kirsten, A., Dengler, P., Ettle, R., & Beer, T. (1996). Selbstbestimmt motiviertes und interessiertes Lernen in der kaufmännischen Erstausbildung [self-determined motivated and interested learning in vocational training in business administration]. *Zeitschrift für Berufs- und Wirtschaftspädagogik*, *13*, 108-127.

Rakoczy, K. (2008). *Motivationsunterstützung im Mathematikunterricht: Unterricht aus der Perspektive von Lernenden und Beobachtern* [Motivational support in mathematics lessons: Instruction from the perspectives of learners and observers]. Münster, Germany: Waxmann.

Rakoczy, K., Buff, A., & Lipowsky, F. (2005). Befragungsinstrumente. [Questionnaires]. In E. Klieme; C. Pauli; K. Reusser (Eds.), *Dokumentation der Erhebungs- und Auswertungsinstrumente zur schweizerisch- deutschen Videostudie „Unterrichtsqualität, Lernverhalten und mathematisches Verständnis", Vol.13.* Frankfurt am Main, Germany: GFPF.

Rakoczy, K., Harks, B., Klieme, E., Blum, W., & Hochweber, J. (2013). Written feedback in mathematics: Mediated by students' perception, moderated by goal orientation. *Learning and Instruction, 27,* 63-73.

Rakoczy, K., Klieme, E., Drollinger-Vetter, B., Lipowsky, F., Pauli, C., & Reusser, K. (2007). Structure as a quality feature in mathematics instruction: Cognitive and motivational effects of a structured organisation of the learning environment vs. a structured presentation of learning content. In M. Prenzel (Ed.), *Studies on the educational quality of schools. The final report on the DFG Priority Programme.* (pp. 102-121). Münster: Waxmann.

Rakoczy, K., Klieme, E., Leiss, D., & Blum, W. (2017). Formative Assessment in Mathematics Instruction: Theoretical Considerations and Empirical Results of the

Co²CA Project. In D. Leutner, J. Fleischer, J. Grünkorn & E. Klieme (Eds). *Competence Assessment in Education: Research, Models, and Instruments* (pp. 447-467). Berlin: Springer.

Ryan, R. M., & Deci, E. L. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary Educational Psychology 25,* 54-67. doi:10.1006/ceps.1999.1020

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science 18,* 119-144.

Schneider, M. C., & Randel, B. (2010). Research on characteristics of effective professional development programs for enhancing educators' skills in formative assessment. In H. L. Andrade & G. J. Cizek (Eds.), *Handbook of Formative Assessment* (pp. 251−276). New York, NY: Routledge.

Seidel, T., & Shavelson, R. J. (2007). Teaching effectiveness research in the past decade: The role of theory and research design in disentangling meta-analysis results. *Review of Educational Research 77*(4), 454-499. doi:10.3102/0034654307310317

Shavelson, R. J., Young, D. B., Ayala, C. C., Brandon, P. R., Furtak, E. M., Ruiz-Primo, M. A., … Yin, Y. (2008). On the impact of curriculum-embedded formative assessment on learning: A collaboration between curriculum and assessment developers. *Applied Measurement in Education, 21,* 295–314. doi:10.1080/08957340802347647

Shute, V. (2008). Focus on formative feedback. *Review of Educational Research*, *78* (1), 153-189. doi:10.3102/0034654307313795

Taylor, B. A., & Fraser, B. J. (2013). Relationships between learning environment and mathematics anxiety. *Learning Environments Research*, *16* (2), 297–313. doi:10.1007/s10984-013-9134-x

TIMSS+ im Nationalfonds-Projekt „Schule, Leistung und Persönlichkeit" (1995). *Population 2. Zweiter Schülerfragebogen M* [Population 2. Second questionnaire]. Bern: Amt für Bildungsforschung.

Von Glasersfeld, E. (1995). *Radical constructivism: A way of knowing and learning*. London: The Falmer Press.

Waldis, M., Buff, A., Pauli, C. & Reusser, K. (2002). *Skalendokumentation zur Schülerinnen- und Schülerbefragung im schweizerischen Videoprojekt* [Scale documentation of the student questionnaire of the Swiss video project]. Universität Zürich: Pädagogisches Institut.

Wang, M. C., Haertel, G. D., & Walberg, H. J. (1993). Toward a knowledge base for school Learning. *Review of Educational Research, 63*(3)*,* 249-294.

Wiliam, D. (2010). An Integrative Summary of the Research Literature and Implications for a New Theory of Formative Assessment. In H. L. Andrade & G. J. Cizek (Eds.), *Handbook of Formative Assessment* (pp. 18−40). New York, NY: Routledge.

Wiliam, D., & Thompson, M. (2008). Integrating assessment with learning: What will it take to make it work?. In C. A. Dwyer (Ed.), The future of assessment. Shaping teaching and learning, (pp. 53–84). New York: Lawrence Erlbaum Associates.

Figures

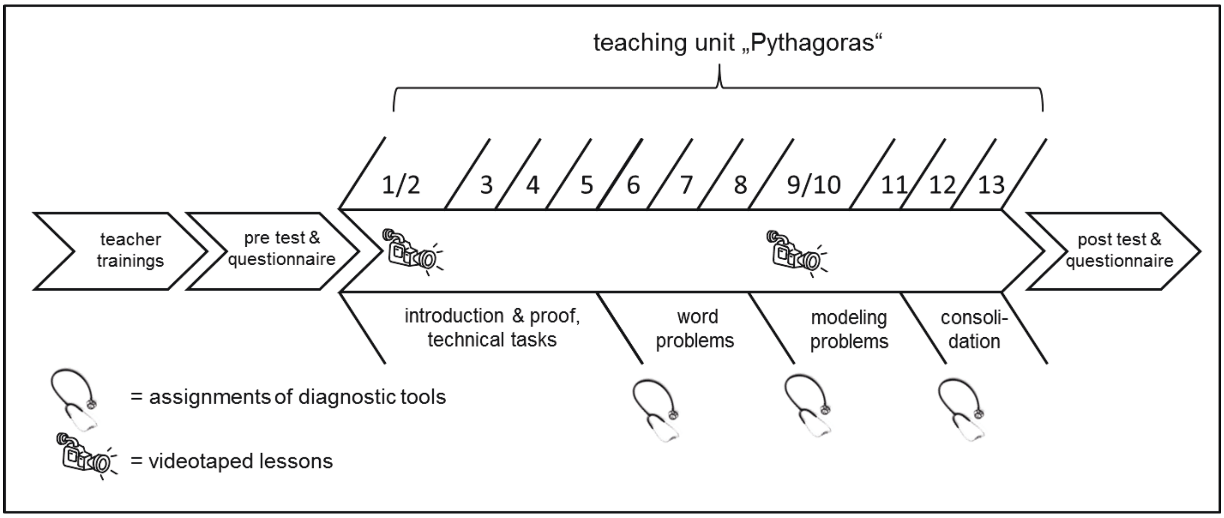

Figure 1. Study Design.

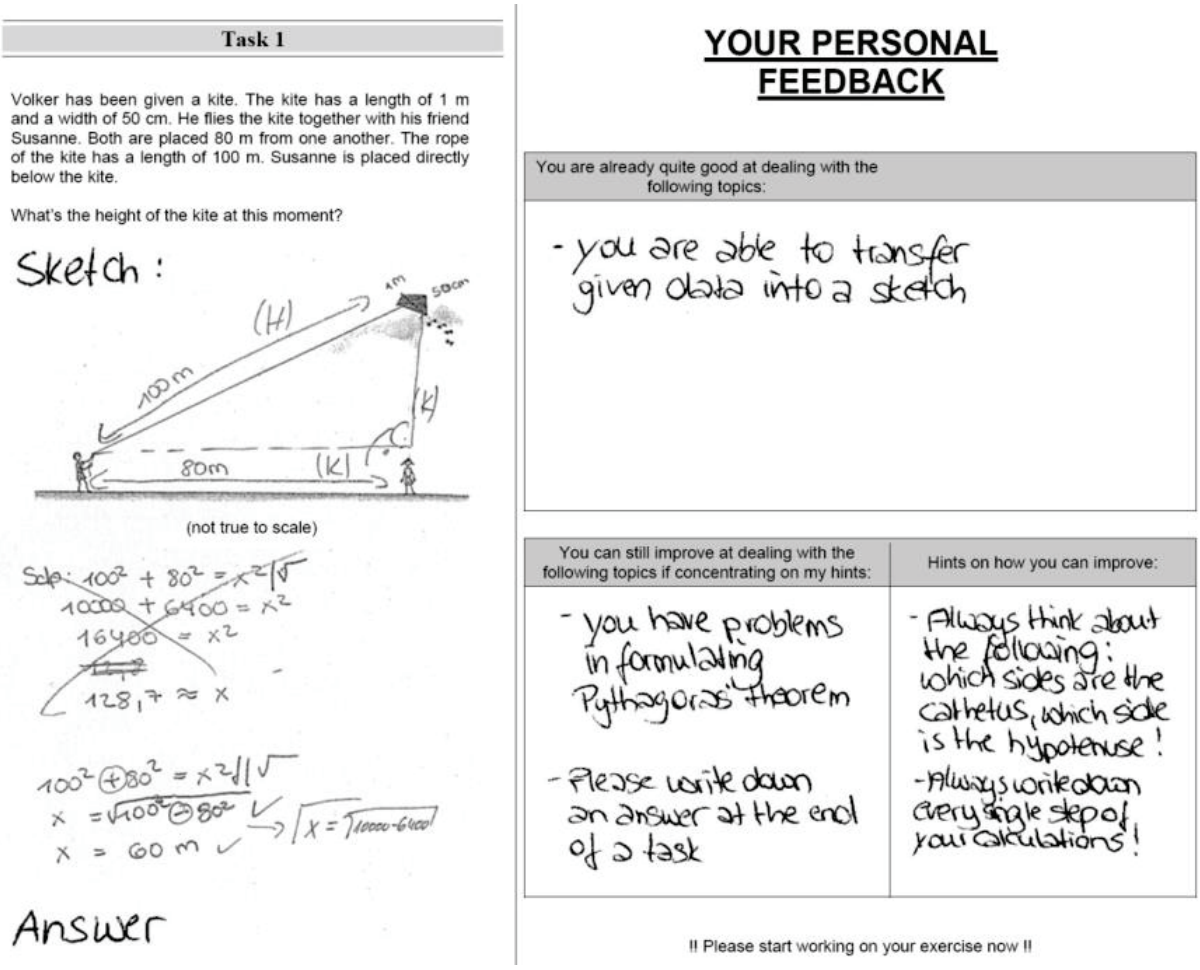

Figure 2. Example of the second diagnostic tool.