

Pinger, Petra; Rakoczy, Katrin; Besser, Michael; Klieme, Eckhard  
**Implementation of formative assessment - effects of quality of programme delivery on students' mathematics achievement and interest**

*formal und inhaltlich überarbeitete Version der Originalveröffentlichung in:*

*formally and content revised edition of the original source in:*

*Assessment in education 25 (2018) 2, S. 160-182*



Bitte verwenden Sie beim Zitieren folgende URN /

Please use the following URN for citation:

urn:nbn:de:0111-pedocs-174067

<http://nbn-resolving.de/urn:nbn:de:0111-pedocs-174067>

#### Nutzungsbedingungen

Gewährt wird ein nicht exklusives, nicht übertragbares, persönliches und beschränktes Recht auf Nutzung dieses Dokuments. Dieses Dokument ist ausschließlich für den persönlichen, nicht-kommerziellen Gebrauch bestimmt. Die Nutzung stellt keine Übertragung des Eigentumsrechts an diesem Dokument dar und gilt vorbehaltlich der folgenden Einschränkungen: Auf sämtlichen Kopien dieses Dokuments müssen alle Urheberrechtshinweise und sonstigen Hinweise auf gesetzlichen Schutz beibehalten werden. Sie dürfen dieses Dokument nicht in irgendeiner Weise abändern, noch dürfen Sie dieses Dokument für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, aufführen, vertreiben oder anderweitig nutzen.

Mit der Verwendung dieses Dokuments erkennen Sie die Nutzungsbedingungen an.

#### Terms of use

We grant a non-exclusive, non-transferable, individual and limited right to using this document.

This document is solely intended for your personal, non-commercial use. Use of this document does not include any transfer of property rights and it is conditional to the following limitations: All of the copies of this documents must retain all copyright information and other information regarding legal protection. You are not allowed to alter this document in any way, to copy it for public or commercial purposes, to exhibit the document in public, to perform, distribute or otherwise use the document in public.

By using this particular document, you accept the above-stated conditions of use.

#### Kontakt / Contact:

peDOCS

DIPF | Leibniz-Institut für Bildungsforschung und Bildungsinformation

Informationszentrum (IZ) Bildung

E-Mail: [pedocs@dipf.de](mailto:pedocs@dipf.de)

Internet: [www.pedocs.de](http://www.pedocs.de)

Mitglied der

  
Leibniz-Gemeinschaft

This is an Accepted Manuscript of an article published by Taylor & Francis in *Assessment in education* on 08/03/2016, available online: <http://www.tandfonline.com/10.1080/0969594X.2016.1170665>.

# **Implementation of formative assessment – effects of quality of programme delivery on students' mathematics achievement and interest**

Petra Pinger, department of educational Quality and evaluation, German institute for international educational research (diPF), Frankfurt, Germany

Katrin Rakoczy, department of educational Quality and evaluation, German institute for international educational research (diPF), Frankfurt, Germany

Michael Besser, Faculty of mathematics, science and technology, university of education Freiburg, Freiburg, Germany

Eckhard Klieme, department of educational Quality and evaluation, German institute for international educational research (diPF), Frankfurt, Germany

## **Abstract**

The aim of this study was to contribute to the understanding of the effectiveness of formative assessment interventions by analysing how the quality of programme delivery affects students' mathematics achievement and interest. Teachers ( $n = 17$ ) implemented formative assessment in their ninth-grade mathematics classes and provided their students ( $n = 426$ ) with written process-oriented feedback. Four feedback characteristics (number of feedback comments, specificity, feedback at self level, social reference norm) and two types of embedment of feedback in the instructional context (focus on feedback utilisation, focus on performance evaluation) were evaluated. Multilevel regression analyses revealed no significant effects of feedback characteristics on interest but negative effects of number of feedback comments and specificity on achievement in mathematics. Positive effects on mathematics achievement and interest were found when feedback was embedded in instruction and had emphasis on feedback utilisation. Students' interest also was affected positively when performance evaluation was stressed.

**Keywords:** Formative assessment; feedback; mathematics; implementation; achievement

## **Introduction**

In educational research, formative assessment is known to be a promising teaching practice in which information on students' understanding is used by means of feedback to promote teaching and learning processes (e.g. Black & Wiliam, 1998). The effectiveness of teaching methods usually is evaluated by measuring students' performance outcomes. A large body of research acknowledges the general positive effects of formative assessment interventions on achievement (Black & Wiliam, 1998; Kingston & Nash, 2011); however, positive effects on students' learning encompass not only achievement but also motivation-related variables. One important construct within the broad range of motivation-related variables is interest. Individuals' interest is predictive of their level of motivation to learn, use of learning strategies and academic achievement (Krapp & Prenzel, 2011, p. 42). Because of its positive long-term effects on learning and educational choices, interest should be seen as an important motivational condition as well as a desired outcome of education (e.g. Krapp & Prenzel, 2011; Kunter, 2005). Mediated through perceived usefulness and perceived competence support, formative feedback has been shown to evoke positive effects on both interest and achievement (Harks, Rakoczy, Hattie, Besser, & Klieme, 2014; Rakoczy, Harks, Klieme, Blum, & Hochweber, 2013).

In general, the effectiveness of formative assessment depends on how formative assessment is realised (Kingston & Nash, 2011) and on the quality of implementation (Furtak et al., 2008). In research on the quality of implementation of formative assessment, adherence to the prescribed intervention and quality of programme delivery have been discussed (e.g. Furtak et al., 2008; O'Donnell, 2008). Research on programme adherence to an intervention addresses whether an intervention was implemented as prescribed; research on the quality of programme delivery addresses how central elements of the intervention are implemented. In their study connecting fidelity of implementation to student achievement, Furtak et al. (2008) evaluated the implementation of the structure of a treatment (adherence) and the process of treatment implementation (quality of delivery). Their results indicated that the quality of delivery was particularly important to the effectiveness of formative assessment.

The aim of this study was to contribute to the understanding of the effectiveness of formative assessment by analysing how aspects of the quality of programme delivery affect both mathematics achievement and interest. In the following, the teaching method of formative assessment, its realisation in the present study and its effects on achievement and interest are described. Subsequently, factors that potentially influence the effectiveness (feedback characteristics and embedment of feedback in instructional context) are explained.

## **Formative assessment**

Formative assessment can be understood as a process that comprises elicitation of information on students' understanding and utilisation of this information to improve students' learning (Black & Wiliam, 1998, 2009; Wiliam & Thompson, 2008). Information can be elicited by 'all those activities undertaken by teachers, and/or by their students which provide information to be used as feedback to modify the teaching and learning activities' (Black & Wiliam, 1998, p. 7). 'All those activities' may include teacher-directed assessment, self-assessment, and peer-assessment, oral and written

assignments and assessments with varying degrees of formality ranging from spontaneous classroom questioning to more formal curriculum-aligned assignment (Shavelson et al., 2008). The utilisation of assessment information can be accomplished in various ways. Teachers can adapt instruction according to the students' level of understanding or use the information to provide students with feedback. In all cases, the intention to affect students' learning positively is implied in the definition of formative assessment, but how well does formative assessment fulfil its purpose of improving students' learning? In a recent meta-analysis, Kingston and Nash (2011) used 42 independent effect sizes derived from 13 studies of formative assessment to calculate a weighted mean effect size of  $d = .20$  on students' achievement (95% confidence interval of .19–.21). Although this might be seen as a realistic estimation of the general effectiveness of formative assessment, it is arguable whether summarising effect sizes across studies of formative assessment provides any meaningful results given that formative assessment is realised in considerably different ways across studies (Bennett, 2011; Black & Wiliam, 1998). Kingston and Nash (2011) therefore concluded that 'research on formative assessment should include clear descriptions of the form and key features of the formative assessment' (p. 35) and that 'research should move from looking at the efficacy of formative assessment to determination of the factors influencing the efficacy of formative assessment' (p. 35). Following that advice, we first describe how formative assessment is realised in the present study and then discuss which aspects of the quality of implementation are investigated to gain insight into the factors influencing the effectiveness of formative assessment.

### **Realisation of formative assessment**

Designing effective formative assessments can be challenging and time-consuming for teachers (e.g. Lee, Feldman, & Beatty, 2011). Predesigned formative assessment material can assist teachers in the realisation of formative assessment, increasing the likelihood of implementation (Hondrich, Hertel, Adl-Amini, & Klieme, 2015). In this study, we implemented curriculum-embedded formative assessment in ninth-grade mathematics instruction. Curriculum-embedded assessments are assignments designed as part of a teaching unit to check students' understanding at critical junctures within the teaching unit. When constructed deliberately, 'formal embedded assessments provide thoughtful, curriculum-aligned, and valid ways of determining what students know' (Shavelson et al., 2008, p. 301). In this study, the assessments were aligned with the secondary school mathematics teaching unit on Pythagoras' theorem. The teaching unit was predesigned by the research team and aimed to foster competency-oriented teaching by focusing not only on technical competencies but also on modelling competencies (as a central part of modern mathematics, see e.g. Niss, 2003 and Bloomhoj & Jensen, 2007). In contrast to working on purely inner mathematical tasks (entirely within mathematics and with no connection to reality), mathematical modelling requires using mathematics to solve real-world problems (Maaß, 2010). The modelling cycle was applied (Blum, Galbraith, Henn, & Niss, 2002; Leiss, Schukajlow, Blum, Messner, & Pekrun, 2010; Maaß, 2006) to develop the teaching unit and interpret students' solutions to the mathematical problems. The modelling cycle is an idealised description of theoretical solution processes, namely cognitive processes, used while performing modelling tasks: first a real-world problem has to be understood, simplified and structured to be transformed into a real model. Second, the real model is mathematicised, resulting in a mathematical model. Third, mathematical methods are applied to calculate a mathematical result. Finally, this result is interpreted and validated with regard to the original real-world problem. If the outcome of the validation process reveals an unsatisfying solution to the problem, the

modelling cycle is restarted. Because solving modelling problems by means of Pythagoras' theorem is challenging, the teaching unit comprised successive phases. During the first phase, Pythagoras' theorem was introduced by a proof and the mathematical procedures of setting up the Pythagoras' theorem and solving equations was practiced by solving technical tasks. During the second phase, embedded word problems were practiced. Like modelling problems, embedded word problems are also set to a context but in a simplified way. In embedded word problems, the real model and all data needed for task solution are given in the text (Maaß, 2010). In the third phase, modelling problems were practiced. During the fourth and final phase, the learning content was consolidated (for an example of an embedded word problem, see Figure 1 and for examples of a technical task and a modelling task, see Appendix). Overall, the teaching unit consisted of 13 lessons distributed over approximately three weeks.

Task 1		YOUR PERSONAL FEEDBACK	
<p>Volker has been given a kite. The kite has a length of 1 m and a width of 50 cm. He flies the kite together with his friend Susanne. Both are placed 80 m from one another. The rope of the kite has a length of 100 m. Susanne is placed directly below the kite.</p> <p>What's the height of the kite at this moment?</p> <p>Sketch:</p> <p>(not true to scale)</p> <p>Sol: <math>100^2 + 80^2 = x^2 \sqrt{}</math>  <math>10000 + 6400 = x^2</math>  <math>16400 = x^2</math>  <math>\sqrt{16400} = x</math>  <math>128,17 \approx x</math></p> <p><math>100^2 + 80^2 = x^2 \sqrt{}</math>  <math>x = \sqrt{100^2 + 80^2} \checkmark \rightarrow x = \sqrt{10000 + 6400}</math>  <math>x = 60 \text{ m} \checkmark</math></p> <p>Answer</p>		<p>You are already quite good at dealing with the following topics:</p> <ul style="list-style-type: none"> <li>- you are able to transfer given data into a sketch</li> </ul>	
<p>You can still improve at dealing with the following topics if concentrating on my hints:</p> <ul style="list-style-type: none"> <li>- you have problems in formulating Pythagoras' theorem</li> <li>- Please write down an answer at the end of a task</li> </ul>		<p>Hints on how you can improve:</p> <ul style="list-style-type: none"> <li>- Always think about the following: which sides are the cathetus, which side is the hypotenuse!</li> <li>- Always write down every single step of your calculations!</li> </ul>	

!! Please start working on your exercise now !!

**Figure 1.** second diagnostic tool (embedded word problem).

The students' understanding was assessed at three critical junctures within the teaching unit (after phases 1–3), each time followed by the provision of process-oriented feedback. Process-oriented feedback is written feedback based on the solution to one or more mathematical tasks and problems. 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; for a detailed description of the formative assessment tool, see Method section). We implemented curriculum-embedded formative assessment and process-oriented feedback in ninth-grade mathematics instruction on Pythagoras' theorem because it allowed us to draw on earlier findings on the effects of process-oriented feedback on achievement and interest. Under laboratory conditions, process-oriented feedback provoked positive effects on achievement and interest on an indirect path via perceived usefulness and perceived competence support (Harks et al., 2014; Rakoczy et al., 2013).

## **Process-oriented feedback**

Feedback generally is assumed to be a powerful tool in initiating student learning processes (e.g. Hattie & Timperley, 2007); however, not all types of feedback are equally effective and – even more important – not every effect of feedback is positive (e.g. Kluger & DeNisi, 1996). In the literature on formative assessment and feedback, there is broad consensus that the effects on students' learning are not reached automatically but that students must understand, accept, and actively process the information provided through assessment and feedback (e.g. Bangert-Drowns, Kulik, Kulik, & Morgan, 1991; Hattie & Timperley, 2007). The characteristics of feedback as well as contextual factors, characteristics of tasks and student variables all influence the way learners process messages transmitted through feedback, and they all influence the direction and magnitude of feedback effects (Bangert-Drowns et al., 1991; Hattie & Timperley, 2007; Kluger & DeNisi, 1996; Shute, 2008). In this study, we implemented one type of feedback developed for mathematics instruction that has been shown to affect learning processes positively: process-oriented feedback (Harks et al., 2014; Rakoczy et al., 2013). The design of process-oriented feedback combines several characteristics considered in the literature on feedback to support cognitive and motivational learning processes. First, process-oriented feedback is elaborate. According to the model of feedback by Hattie and Timperley (2007), effective feedback should provide a sufficient amount of information to answer the three questions: Where am I going? How am I going? and Where to next? (Hattie & Timperley, 2007). Process-oriented feedback comprises this information by informing the learner about his or her strengths and weaknesses, and providing him or her with strategies to close the gap between the learning goal(s) and his or her current level of understanding.

Second, feedback can be provided at task level (information on task performance), process level (information on processes required to master the task), self-regulatory level (information on the regulation of action) and self level (information on the learner as a person, not related to task performance; Hattie & Timperley, 2007). While feedback at the first three levels is associated with positive learning outcomes, feedback at the self level usually contains too little task-related information to show positive effects on learning processes (Hattie & Timperley, 2007). The design of process-oriented feedback draws on Hattie and Timperley's (2007) ideas and tries to combine feedback at task level, process level and self-regulatory level by relating feedback to concrete tasks and by focusing on cognitive and self-regulatory processes (Harks et al., 2014; Rakoczy et al., 2013).

Third, process-oriented feedback supports the use of individual-reference norms and criterion-reference norms by focusing on the individual performance profile (strengths and weaknesses) and processes needed to perform the tasks (strategies; Harks et al., 2014; Rakoczy et al., 2013). An individual-reference norm can be used in feedback to provide intra-individual comparisons focusing on the individual's development or profile, and a criterion-reference norm can be used to evaluate performance against absolute standards. A social reference norm can be used to compare an individual's achievement to the performance of other students in the class (e.g. Krampen, 1987; Rheinberg, 2006). Empirical findings have shown more beneficial effects of using individual and criterion frames of reference than using social reference norms (e.g. Harks et al., 2014; Krampen, 1987; Rakoczy et al., 2013; Rheinberg, 2006; Shih & Alexander, 2000).



### **Effects of process-oriented feedback on achievement**

Process-oriented feedback is supposed to support achievement development by providing task-specific and process-specific information on how to overcome discrepancies between learning goals and current understanding. Information on whether or not the task was completed correctly helps the learner identify misconceptions and facilitates error correction. Additional information on strengths, weaknesses and strategies provides learners with appropriate problem-solving strategies and encourages them to take responsibility for their own learning. It has a self-efficacy enhancing function and guides learners in the next steps in the learning process (Harks et al., 2014; Rakoczy et al., 2013).

### **Effects of process-oriented feedback on interest**

Interest is the motivational relationship between a person and subject matter (object) and therefore can be said to be a content-specific type of motivation. According to the person-object approach to interest (POI), developing and maintaining interest depends on cognitive-rational and emotional components (e.g. Krapp, 1999, 2002, 2005). The cognitive-rational or value component of this approach states that a person will be interested in learning content that has personal significance. The emotional component refers to positive feelings associated with engagement in that subject matter. As interest is linked theoretically to the concept of intrinsic motivation, the theoretical framework of the POI draws on self-determination theory (e.g. Ryan & Deci, 2000). Self-determination theory postulates that supporting the three basic needs for competency, autonomy and social relatedness is relevant to the emotional component of developing interest. Process-oriented feedback provides information on competencies that have been mastered and on how to improve. This information is meant to fulfil the learner's basic need to feel competent as well as to enhance the learner's self-efficacy in performing a task in a specific content area. This theoretical link has been supported by results from the laboratory studies conducted by Rakoczy and colleagues (2013) and Harks and colleagues (2014), in which interest development was supported indirectly through perceived usefulness and perceived competence support.

### **Quality of programme delivery**

Because of its formative function and its theoretical and empirical links to mathematics achievement and interest development (Harks et al., 2014; Rakoczy et al., 2013), process-oriented feedback should be a promising realisation of formative assessment. However, results of some studies have indicated that the quality of programme delivery might be of particular importance concerning the effectiveness of formative assessment (e.g. Furtak et al., 2008). As noted above, quality of programme delivery refers to the way in which key elements of an intervention are implemented. In this study, we are interested in how two aspects of the quality of delivery of process-oriented feedback affect achievement and interest: (i) the process of generating feedback, that is, characteristics of the written feedback, and (ii) the process of relaying feedback to the students, that is, how the formative assessment tool is embedded in the instructional context.

## Characteristics of written feedback

Due to the design of our study and of the design of process-oriented feedback, some parts of the intervention were held constant across the participating classes in terms of the number of times feedback was given (three times), the assessment tasks, the timing of feedback (provided the next lesson) and the general structure of feedback (strengths, weaknesses and strategies were preset; see Method section). Four characteristics of the feedback were not predetermined and therefore might have varied and affected students' achievement and interest in the mathematical topic: number of feedback comments, feedback at self level, specificity of feedback and the use of a social reference norm.

**Number of feedback comments.** The effects of the elaborateness of feedback often are studied in terms of type of feedback, which varies in quantity of content. For example, feedback that provides information only on the correctness of a response contains less information than feedback offering explanations or strategies to improve (e.g. Shute, 2008). In general one might assume that more elaborate feedback will have greater effects, but results of research comparing the effectiveness of simple feedback to that of more elaborate types of feedback are not consistent (e.g. Kulhavy, White, Topp, Chan, & Adams, 1985). Discussing these inconsistent results, Huth (2004) explains that less informative feedback might be sufficient for tasks in which only declarative knowledge is needed while more elaborate feedback could be more effective for tasks requiring procedural knowledge. In this regard, elaborate feedback has positive effects on motivation and achievement when completing arithmetic tasks requiring procedural and conceptual knowledge (Narciss & Huth, 2006). Process-oriented feedback is by design an elaborate type of feedback because it provides information not only on the correctness of a response but also on strengths, weaknesses and strategies to improve. Thus, the structure of the feedback is predesigned, but the number of feedback comments and content of feedback the learners receive can vary. The number of feedback comments refers not to the amount of sentences but rather on the number of competencies on which feedback is given. As solving embedded word problems requires multiple procedural steps and sub-competencies, we assumed that giving students detailed feedback on the competencies they had mastered and strategies to develop sub-competencies would lead to better planning of the next steps in learning, an increased sense of competency and self-efficacy, and consequently, to positive effects on achievement and interest.

**Feedback at self level.** Process-oriented feedback is designed to direct the learner's attention to and provide information on the processes underlying the assessment tasks. With regard to the four feedback levels (Hattie & Timperley, 2007), process-oriented feedback is provided at the task level (informing the student whether or not an assessment task was completed correctly) and at the cognitive and self-regulation processes level (informing the student about his or her strengths and weaknesses, and suggesting strategies to improve, ideally referring to processes and operations needed to complete tasks). Although process-oriented feedback is designed to avoid feedback at self level, feedback can be provided at this level (in written statements about the learner as a person, e.g. 'you are a math wizard'). Drawing on the literature, Hattie and Timperley (2007) claim that praise as one type of self-level feedback (when directed to the learner as a person and not to a task or

processes) has little to no effect on achievement. Furthermore, studies of the effects of praise indicate that self-level praise can have a counter-intuitive negative effect on perceived ability and consequently on self-efficacy (Meyer, 1982). In summary, Hattie and Timperley (2007) conclude that feedback at the self level 'contains little task-related information and is rarely converted into more engagement, commitment to the learning goals, enhanced self-efficacy, or understanding about the task' (p. 96).

**Specificity of feedback.** The general structure of process-oriented feedback (strengths, weaknesses and strategies to complete tasks and develop the necessary underlying processes to do so) is predefined. The degree of specificity may vary from global feedback statements (e.g. 'you should study Pythagoras' theorem again') to specific feedback comments (e.g. 'you still have problems identifying the hypotenuse in a right-angled triangle'). Unspecific feedback can lead to uncertainty about how to respond, to the perception of the feedback as useless and to higher cognitive load which in turn can lead to reduced learning and motivation (Shute, 2008, pp. 157–158). This may hold true especially for feedback concerning weaknesses and strategies. Unspecific feedback on weaknesses could direct attention to the self and threaten the learner's self-esteem, and unspecific strategies might leave the learner uncertain as to how to proceed in the learning process.

**Social reference norm.** Feedback comments can be characterised by the reference norm orientation applied (Rheinberg, 2006). As noted above, feedback can be provided using a social reference norm to compare a student's achievement to that of other students in the class. The use of social reference norms by grading and reporting the grade average of a class is common in everyday classroom practices and is associated with summative evaluations rather than with formative feedback (Ames, 1992; Cizek, Fitzgerald, & Rachor, 1996; Köller, 2005). Empirical findings indicate that criterion-related and individual reference norms are more advantageous for low achieving students (e.g. Krampen, 1987). As formative feedback such as process-oriented feedback aims to track and support individual learning progress concerning particular criteria or learning goals, the use of criterion-related and individual-reference norms is advisable and the use of a social reference norm should be avoided (for an overview of empirical findings on formative assessment and reference norms, see Köller, 2005).

### **Embedment of feedback in instructional context**

A feedback message will affect learning processes positively only if information provided in the feedback is perceived and actively processed by the learners (Bangert-Drowns et al., 1991). Therefore, attention must be directed to the relevant information provided in the feedback. How students perceive, process and make use of process-oriented feedback might be influenced by the context in which it is given. Support for the notion that context influences the way feedback is perceived and processed comes from a study conducted by Cianci, Schaubroeck, and McGill (2010). In their study, direction and magnitude of feedback effects on achievement was influenced by the context in which individuals received feedback. The performance of participants receiving learning goal instructions improved to a greater extent when they received negative feedback than when they received positive feedback; the opposite pattern was observed among participants who had received performance goal instructions. Although results of this study cannot be transferred directly to the

present study, the findings indicate that the context in which assessment and feedback is provided influences its effect.

The way teachers introduce the provision of feedback to their students in the classroom is assumed to influence the way it is perceived and processed. This is what we refer to as embedment of feedback. It might direct students' attention and alter feedback effects. Feedback can be presented to students by stressing performance evaluation, directing their attention to the correctness of responses (e.g. 'the test results were satisfactory') or by stressing the utilisation of feedback, directing their attention to relevant feedback information (e.g. 'read the feedback comments carefully').

### **Research aim, questions and hypotheses**

Drawing on the theoretical and empirical findings in the literature on formative assessment and feedback discussed in the previous sections, we identified feedback characteristics and the embedment of feedback in instruction as aspects of the quality of delivery of the formative assessment intervention. The aim of this study was to contribute to the understanding of the effectiveness of formative assessment by evaluating how these factors affect mathematics achievement and interest.

- (1) How do teachers provide process-oriented feedback in terms of number of feedback comments, specificity, self level and social reference norm?

By design, process-oriented feedback is provided at the task level and process level and focuses the use of individual-reference and criterion-reference norms. Therefore, we expected teachers not to provide feedback at the self level and not to use a social reference norm. The number of feedback comments and the specificity of feedback were not predesigned; therefore, interindividual variation was expected.

- (2) How do feedback characteristics affect students' mathematics achievement and interest?

Given that the more feedback comments provided, the more relevant information should be provided, we expected a positive effect of the number of feedback comments on mathematics achievement (2a) and interest (2b). We also expected a positive effect of highly specified feedback on mathematics achievement (2c) and interest (2d).

- (3) How does the way in which feedback is embedded in the instructional context affect mathematics achievement and interest?

Encouraging feedback utilisation by emphasising the usefulness of the information provided in the feedback should direct students' attention to the informative elements of the feedback and lead to positive effects on students' mathematics achievement (3a) and interest (3b). Stressing the evaluation of the responses should direct students' attention more to the correctness of the response and less to the additional information provided in the feedback. Therefore, we expected no effect of the embedment of feedback as performance evaluation on achievement (3c) or interest (3d).

## **Method**

The intervention was conducted in the academic year 2010/2011 as part of the project 'Conditions and Consequences of Classroom Assessment (Co<sup>2</sup>CA)' 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.<sup>1</sup> 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 (in press). The study of the quality of delivery of the written formative assessment presented here is part of a larger quasi-experimental study investigating the impact of formative assessment interventions on learning. The intervention groups implementing written curriculum-embedded formative assessments were included in the present analyses.

## **Participants**

The 24 teachers from 14 middle track state schools in Hesse, Germany (urban and rural areas) participating in this intervention study were recruited by inviting them to attend a professional development workshop on assessment and feedback. Participation in the study was voluntary for teachers and students. Written informed consent of the parents was obtained for all students participating in the study and all data protection requirements were met.

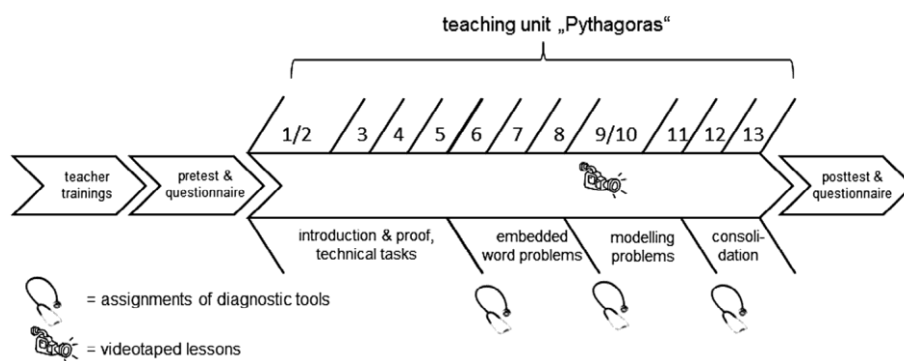
To avoid confounding results of the effects of the quality of delivery with potential effects of frequency and timing of feedback, we excluded all classes in which one or more diagnostic tools had not been assigned or had not been returned to students in the subsequent lesson. Seven teachers were excluded because of drop out caused by illness ( $n = 2$ ), because not all three diagnostic tools were assigned ( $n = 1$ ) or because feedback was not provided in the subsequent lesson ( $n = 4$ ). The final subsample consisted of 17 teachers and 426 students (46.6% female; mean age 15.4 years). The teachers had on average 16.2 years of working experience and 62.5% were female.

## **Design and procedure**

The intervention was explained to the teachers during two half-day training sessions. During the first training session, teachers were provided with organisational information and subject-specific content concerning the predesigned teaching unit on Pythagoras' theorem. The teaching unit had four phases: (1) an introduction including a proof and technical tasks, (2) embedded 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. During the second training session, teachers were instructed on how to implement the diagnostic tools and provide process-oriented feedback. At the end of each phase of the teaching unit (at the end of lessons 5, 8 and 11), teachers were asked to administer a diagnostic tool to assess students' understanding and to provide process-oriented feedback. After the first phase, two technical tasks were assigned; after the second, an embedded word problem was assigned; and after the third, a modelling problem was

assigned. 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).

To assist teachers in the formulation of feedback and to ensure quality of inference drawn from the students' solutions, a semi-structured feedback procedure was developed. Cognitive task analyses revealed a list of cognitive processes and operations needed to complete the diagnostic tasks. If a particular cognitive process was mastered, it was considered a strength and teachers gave feedback as such. If a process had not been mastered, it was considered a weakness, teachers gave feedback as such, and provided a corresponding recommendation for improvement. To prevent long and



**Figure 2.** study design.

complex feedback, sub-competencies could be subsumed under main competencies. For example, the sub-competencies 'identifying cathetus and hypotenuse' and 'formulating Pythagoras' theorem correctly' could be subsumed under the main competency 'making use of Pythagoras' theorem in embedded word problems'. Teachers were allowed to give additional feedback that was not recommended on the list. Concerning feedback levels and reference norms, teachers were instructed not to use feedback at the self level and to avoid the use of social reference norms. Teachers returned the corrected diagnostic tool with process-oriented feedback during the subsequent lesson (beginning of lessons 6, 9 and 12). The procedure of providing feedback was specified in the intervention guidelines. The teachers were instructed first to hand back the corrected diagnostic tool with written feedback and to give the students enough time to read them. Then, the teachers were to administer another problem very similar to the diagnostic task. To gain insight into how the teachers gave the feedback to students, lessons 9 and 10 (returning second corrected diagnostic tool and feedback) were videotaped.

Prior knowledge was assessed immediately before the intervention and achievement in mathematics was assessed immediately afterwards. Students' interest in the topic of the test was assessed using questionnaires prior to the pretest and the posttest (see Figure 2 for an overview of the study design).

## **Measures**

### **Achievement**

In the lesson before beginning, the predesigned teaching unit students' prior knowledge was assessed using a pretest (19 items); in the lesson immediately after it, mathematical achievement on Pythagoras' theorem was assessed using a posttest (17 items). The pretest did not assess knowledge of Pythagoras' theorem; rather it assessed relevant prior knowledge such as identifying a right-angled triangle and solving equations. The posttest included technical tasks, word problems and modelling tasks (for examples of pretest and posttest items, see the Appendix). The items had been analysed previously in a scaling study ( $N = 1570$ ); therefore, item parameters from the one-dimensional Rasch model of the scaling study could be used as fixed parameters in the one-dimensional Rasch model applied to the quasi-experimental data. Weighted likelihood estimator parameters served as achievement scores for pretests and posttests. Analyses were conducted in ConQuest (Wu, Adams, & Wilson, 1998). Estimated reliability (EAP/PV) was .66 for the pretest and .74 for the posttest.

Students did not receive feedback on their performance on the pretest or posttest but teachers were told how many students in their classes had completed the first three items of the test correctly.

### **Interest**

As individual interest is seen as a slowly alterable construct (e.g. Krapp, 2002), we measured interest specifically related to the topic of the test rather than interest in the general subject of mathematics. Students' interest in the topic of the test was self-reported on a questionnaire immediately before the pretest and again immediately before the posttest (adapted from Rakoczy, Buff, & Lipowsky, 2005). Before responding to the items on the questionnaire, students were asked to look at all the items on the pretest or posttest. By doing so, interest in the topic of the test was related to tasks and problems that were directly linked to our intervention. Students indicated on a four-point scale ranging from 0 (completely disagree) to 3 (completely agree) how interested they were in the topic of the test (e.g. 'I like the topic of the test'). Internal consistency of the scale was Cronbach's  $\alpha$  .83 for the pretest questionnaire and .89 for the posttest questionnaire.

### **Characteristics of the feedback**

To evaluate the characteristics of the feedback, we analysed the written feedback of the second diagnostic tool because it had been administered at a central juncture within the teaching unit and combined technical and modelling competencies. Of the 426 participating students, 378 completed this task and received feedback. The 378 feedback sheets were analysed with regard to the four characteristics of feedback under investigation. To determine the number of feedback comments for each student, comments written by the teacher about his or her strengths, weaknesses and recommendations for improvement were calculated for main competencies and sub-competencies, and any additional comments were tallied. In the example in Figure 1, the student received five

feedback comments: one sub-competency was fed back as strength, two sub-competencies as weaknesses and for two competencies the teacher provided a recommendation for improvement.

To evaluate specificity of feedback, feedback at the self level and use of a social reference norm, a coding scheme was developed. Specificity was judged separately for weaknesses and recommendations, and then averaged. Code 0 was assigned when no comment was specific (e.g. 'you should study Pythagoras' theorem again') and code 1 when at least one comment was specific (e.g. 'check first: which sides are the cathetus and which side is the hypotenuse'). In our example, all feedback comments were formulated specifically.

When feedback was judged to be at self level, code 1 was assigned ('you are a math wizard!') and code 0 was assigned when feedback was judged not to be at self level. For the category social reference norm, code 1 was assigned when the student's individual performance was compared to the performance of the class (e.g. 'compared to most of the students in the class you performed well/poorly'). In our example in Figure 1, no feedback was provided at the self level. Also, no comparison to other students or the performance of the class was made (for examples of unspecific feedback, feedback at the self level, and the use of a social reference norm, see Appendix).

Two raters were trained according to a manual describing the coding system. After two coding rounds, the inter-rater reliability was sufficient with Cohens- $\kappa$  of .97 and 1, respectively.

### **Embedment in instruction**

To evaluate how the diagnostic tool was embedded in instruction, the ninth lesson, in which the second diagnostic tool was returned to the students, was videotaped. One teacher could not be videotaped; therefore, 16 videotaped lessons were available. High-inference ratings were applied to evaluate how teachers embedded the diagnostic tool in their instruction. Two aspects of types of embedment were assessed: embedment as 'feedback utilisation' (code 0 = diagnostic tool was returned without emphasising feedback utilisation; code 1 = utilisation of feedback was emphasised once, e.g. 'read the feedback comments carefully, they can help you to do even better next time'; code 2 = repeated emphasis was placed on feedback utilisation) and embedment of the diagnostic tool as 'performance evaluation' (code 0 = performance evaluation was not stressed in instruction; code 1 = performance evaluation was emphasised; e.g. 'the test result was satisfactory'). Two raters were trained according to a manual describing the codes. Due to an inter-rater agreement of Cohens- $\kappa$  of .28 and .15, respectively, final coding was discussed and reached by consensus.

### **Data analysis**

We conducted multilevel regression analyses using Mplus 7 to account for the nested data structure (Muthén & Muthén, 1998–2012). Before analysing the data in Mplus, scores for mathematics achievement and interest and for the number of feedback comments and specificity of feedback were z-standardised in SPSS. Separate analyses were run for the effects on achievement and interest, as well as for each characteristic of feedback and each type of embedment. The z-standardised pretest score for achievement in mathematics and the z-standardised pre-questionnaire score for



interest as well as z-standardised scores for characteristics of feedback were entered as level 1 predictors. Scores on the embedment in instruction were entered as level 2 predictors.

## Results

Our first aim was to investigate how process-oriented feedback was realised in terms of the number of feedback comments, how specific the feedback was, whether feedback was provided at the self level, and whether it made use of social reference norms. As expected, almost no feedback was directed at the self level and in hardly any case was a social reference norm used; these two characteristics of feedback were observed only once. On average, students received 5.35 feedback comments (SD = 2.44), and the average feedback specificity was relatively high with  $M = .82$  (SD = .30). Tables 1 and 2 present the descriptive data including sample sizes, means, standard deviations and correlations<sup>2</sup> separately for student-level (level 1) and classroom-level (level 2) variables.

Intra-class correlations (ICC) were .045 for posttest achievement and .051 for posttest interest. Although ICCs were relatively low, they indicate that classes differed with regard to the composition of posttest achievement and posttest interest.

**Table 1.** sample sizes, correlations, means and standard deviations (level 1).

Variable	<i>n</i>	1	2	3	4	5	<i>M</i>	<i>SD</i>
1 Achievement pre	374	—					−1.13	.86
2 Achievement post	391	.49**	—				−.15	1.06
3 interest pre	367	.09	.13*	—			2.29	.64
4 interest post	390	−.00	.10	.39**	—		2.62	.71
5 number of comments	376	−.25**	−.22**	.07	.07	—	5.35	2.44
6 specificity	379	−.13*	−.14*	−.07	.05	.10	.82	.30
7 self-level	378	—	—	—	—	—	.003	.05
8 social-reference	378	—	—	—	—	—	.003	.05

\* $p < .05$ , two-tailed.

\*\* $p < .01$ , two-tailed.

**Table 2.** sample sizes, correlations, means and standard deviations (level 2).

Variable	<i>n</i>	1	2	3	4	5	6	<i>M</i>	<i>SD</i>
1 Achievement pre	17	—						−1.12	.32
2 Achievement post	17	.65**	—					−.15	.31
3 interest pre	17	−.13	−.21	—				2.29	.18
4 interest post	17	−.13	.09	.53*	—			2.62	.22
5 Feedback utilisation 1 <sup>a</sup>	16	.21	.60*	−.07	.08	—		.25	.48
6 Feedback utilisation 2 <sup>b</sup>	16	−.32	−.31	−.21	.11	−.39	—	.31	.48
7 summative test <sup>c</sup>	16	−.21	.15	−.39	.26	−.08	.13	.31	.48

<sup>a</sup>dummy-coded: feedback utilisation 1: emphasis on feedback once = 1, else = 0.

<sup>b</sup>dummy-coded: feedback utilisation 2: emphasis on feedback repeatedly = 1, else = 0.

<sup>c</sup>dummy-coded: emphasis on performance evaluation = 1, no emphasis = 0.

\* $p < .05$ , two-tailed.

\*\* $p < .01$ , two-tailed.

**Table 3.** multilevel regression analyses of aspects of 'quality of delivery' on achievement.

	Model 1		Model 2		Model 3		Model 4	
	$\beta$	SE	$\beta$	SE	$\beta$	SE	$\beta$	SE
<i>Individual level</i>								
Achievement pre	.52**	.08	.53**	.08	.49**	.06	.50**	.06
number of comments	-.13*	.06	—	—	—	—	—	—
specificity	—	—	-.10*	.05	—	—	—	—
<i>Class-level</i>								
Feedback utilisation 1 <sup>a</sup>	—	—	—	—	.40**	.14	—	—
Feedback utilisation 2 <sup>b</sup>	—	—	—	—	.10	.13	—	—
Performance evaluation <sup>c</sup>	—	—	—	—	—	—	.14	.16

<sup>a</sup>dummy-coded: feedback utilisation 1: emphasis on feedback once = 1, else = 0.

<sup>b</sup>dummy-coded: feedback utilisation 2: emphasis on feedback repeatedly = 1, else = 0.

<sup>c</sup>dummy-coded: emphasis on performance evaluation = 1, no emphasis = 0.

\* $p < .05$ , two-tailed.

\*\* $p < .01$ , two-tailed.

**Table 4.** multilevel regression analyses of aspects of 'quality of delivery' on interest.

	Model 1		Model 2		Model 3		Model 4	
	$\beta$	SE	$\beta$	SE	$\beta$	SE	$\beta$	SE
<i>Individual level</i>								
interest pre	.25**	.04	.25**	.04	.30**	.03	.41**	.04
number of comments	.03	.04	—	—	—	—	—	—
specificity	—	—	.05	.03	—	—	—	—
<i>Class-level</i>								
Feedback utilisation 1 <sup>a</sup>	—	—	—	—	.15	.10	—	—
Feedback utilisation 2 <sup>b</sup>	—	—	—	—	.18*	.09	—	—
Performance evaluation	—	—	—	—	—	—	.21*	.10

<sup>a</sup>dummy-coded: feedback utilisation 1: emphasis on feedback once = 1, else = 0.

<sup>b</sup>dummy-coded: feedback utilisation 2: emphasis on feedback repeatedly = 1, else = 0.

<sup>c</sup>dummy-coded: emphasis on performance evaluation = 1, no emphasis = 0.

\* $p < .05$ , two-tailed.

\*\* $p < .01$ , two-tailed.

Concerning our second research question, we expected a positive effect of number of feedback comments and specificity of feedback on mathematics achievement (2a, 2c) and interest (2b, 2d). The multilevel analyses of the number of feedback comments and specificity of feedback as level 1 predictors and pretest scores as covariates revealed negative effects on posttest achievement scores. Examination of the correlation table revealed that the number of feedback comments and specificity of feedback correlated negatively with pretest achievement scores ( $r = -.25$ ,  $p < .01$  and  $r = -.13$ ,  $p < .05$ , respectively), indicating that students with low scores on the pretest received more feedback remarks and more specific feedback. Moreover, we found significant negative correlations between the number of feedback comments and specificity of feedback and posttest achievement ( $r = -.22$ ,  $p < .01$  and  $r = -.14$ ,  $p < .05$ , respectively). With regard to post-questionnaire interest in the topic of the test, no significant effects of the number of feedback comments or specificity of feedback were found. Beta-coefficients and standard deviations for the multilevel regression analyses are summarised in Table 3 for achievement and Table 4 for interest.

With regard to our third research question concerning the embedment of the diagnostic tool in instruction, multilevel regression analyses showed a positive effect of a moderate emphasis on feedback utilisation (feedback utilisation 1) on achievement in mathematics, but no effect when emphasis was placed repeatedly (feedback utilisation 2; 3a). Concerning the effects on interest in the

topic of the test, an opposite pattern emerged. Repeated emphasis on the utilisation of feedback (feedback utilisation 2) had a positive effect on students' interest while stressing utilisation only once (feedback utilisation 1) did not (3b). Contrary to our expectations, positive effects on students' interest also were found for the embedment as performance evaluation (3d). As hypothesised, no effect of embedment as performance evaluation on achievement was found (3c).

## Discussion

The aim of the present study was to contribute to the understanding of the effectiveness of formative assessment by analysing how aspects of the quality of its delivery affect mathematics achievement and interest. Concerning our first research question, investigation of the feedback teachers provided revealed that process-oriented feedback was generated as intended by the feedback design. Feedback at the self level and the use of social reference norms were avoided in nearly all cases. This is not self-evident given that grading as a form of social comparative feedback and praise as form of feedback at the self level are rather common in everyday classroom assessments (cf. Hattie & Timperley, 2007). The design of the diagnostic tool and the provision of the semi-standardised feedback procedure probably helped teachers avoid these types of feedback. The headings of the text fields of the diagnostic tool referred to processes and operations that the individual student has (not) successfully mastered. Additionally, teachers' written feedback was guided by the list of cognitive processes needed for task solution and corresponding hints. This list served as very specific examples of feedback comments.

With regard to our second research question, we expected the number of feedback comments and specificity of feedback to have positive effects on achievement and interest. Contrary to our expectations, we found no significant effects on interest and negative effects on posttest achievement. First, the negative effects of the number of feedback comments on posttest achievement in mathematics can be explained partly by the finding that the number of feedback comments was negatively correlated with pretest achievement, indicating that lower achieving students received more feedback comments and that those students possibly did not benefit from the feedback they received regarding their learning progression. Second, we assumed that the more feedback comments that were given, the more relevant information would be included, leading to an increased chance of students correcting their errors and guidance in adapting their learning processes. However, the increased number of feedback comments might have led to greater complexity of the feedback message, and complex feedback does not always lead to positive learning outcomes (for an overview, see Shute, 2008). In a study conducted by Kulhavy et al. (1985) feedback complexity was inversely related to error correction. The authors concluded that learners might have processed the more complex feedback at a shallower level or that complex feedback might have been perceived as less useful (Kulhavy et al., 1985, pp. 290–291).

One possible explanation for the negative effects of specificity on achievement is that feedback can be too specific and thereby not transferable to other tasks (e.g. Hattie & Timperley, 2007). In this study, we analysed the specificity of feedback on one particular type of task, namely to an embedded word problem, while the aim of the whole teaching unit was to understand and apply Pythagoras' theorem in various types of tasks and problems. The posttest, which covered the content of the

entire teaching unit, might not have been sensitive enough to detect positive effects of specific feedback on one type of problem.

Concerning our third research question, we expected the embedment of formative assessment as 'feedback utilisation' to have positive effects on mathematics achievement and interest. Our results revealed a positive effect on achievement when feedback utilisation was emphasised once but not if it was emphasised repeatedly. One explanation for these results could be that emphasising feedback utilisation once (and not repeatedly) is an indicator that providing feedback in class might have become routine by the ninth lesson of the teaching unit. That, in turn, could be interpreted as an indication of a good classroom management (e.g. Klieme, Pauli, & Reusser, 2009). A higher degree of classroom management in turn has been demonstrated to be an important predictor of students' learning (e.g. Hattie, 2009).

Regarding the effects of embedment of formative assessment as 'feedback utilisation' on interest in the topic of the test, an opposite pattern emerged. Emphasising the utilisation of feedback repeatedly had a positive effect on students' interest while stressing utilisation only once did not. Contrary to our expectations, emphasising 'performance evaluation' had positive effects on students' interest. Stressing both performance evaluation and the utilisation of feedback repeatedly might have underlined the relevance of the content of the subject matter. The perception of the relevance of content might have facilitated internalisation of the extrinsically motivated activity, which in turn was associated with more interest (e.g. Krapp, 2002, 2005; Ryan & Deci, 2000).

### **Limitations and implications for further research**

The present study has several limitations that need to be discussed. First, the small sample size of 24 teachers was further reduced due to insufficient adherence to the intervention. Therefore, the generalisability of our results is questionable. To replicate and verify the findings of this study future research should be conducted with larger samples. Second, the posttest of this study consisted of various types of tasks to assess the global understanding of Pythagoras' theorem. The posttest might not have been sensitive enough to detect the effects of the characteristics of the feedback on the second diagnostic tool. Third, we found quite low inter-rater reliability for the ratings of the videotaped recordings of teachers embedding the written feedback into instruction. As we had no comparable videos available to calibrate the raters, we had to do the calibration with the original videotapes of the study. Discrimination between the codes apparently was not clear enough, as rater discussion was needed to reach consensus and to improve the coding scheme over the course of the rating process. Moreover, there are general limitations of the video ratings (Stigler, Gallimore, & Hiebert, 2000). For example, a video sequence always contains only an excerpt of a teacher's classroom behaviour. Teachers might have acted differently each time they employed the assessment tools. Also, teachers might have acted differently while being videotaped from how they usually behave in the classroom because they knew that they were being filmed. Although interpreting – and especially generalising – the findings of video analyses must be done with caution, results of our study indicate the importance of context in which feedback is provided. This is in line with other research findings that reveal, for example, that the general teaching quality has a positive influence on the effectiveness of formative assessment (Decristan et al., 2015).

Lastly, and maybe most importantly, our teacher training and the intervention period were relatively short. Our results indicate that the training enabled those teachers who implemented the formative assessment tool at all three junctures to provide feedback as intended; however, it is questionable as to whether the assessment culture changed during this short period from the usual assessment of learning to an assessment for learning (e.g. Tierney, 2006). Although our assessments had a formative design, a formative intention, and no summative function, they might have been perceived by the students as summative tests. Our finding that some teachers stressed performance evaluation when they provided feedback indicates that the formative function of the assessments might not have been internalised by the teachers or students.

### **Concluding remarks**

In this study, we investigated the effects of various aspects of the quality of programme delivery on achievement and interest in mathematics, focusing on four feedback characteristics and two contextual factors. We found that the participating teachers provided process-oriented feedback as intended by the feedback design and explained during the teacher training. However, an increase in the amount of information in, and the specificity of, feedback comments did not result in the expected positive effects. Because 'mindful processing' is essential for feedback to show effects (e.g. Bangert-Drowns et al., 1991), future research should include the learners' perception of the formative assessment intervention, the learners' intention to make use of the feedback and the learners' actual use of the feedback provided. From previous studies we know about the role of perceived usefulness and perceived competence support as an important mediator in the working mechanisms of feedback processing (Harks et al., 2014; Rakoczy et al., 2013). Instead of directly linking the characteristics of the feedback and the contextual factors to posttest achievement and interest, it might be useful to integrate theoretically assumed mediators.

Although results were in part contrary to our expectations, they reveal that the way formative assessment and in particular process-oriented feedback is generated and delivered, does affect cognitive and motivational processes. Our results underline the importance of further research to be able to provide teachers with advice on how to design and implement formative assessment and feedback effectively.

### **Notes**

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).
2. Because the feedback characteristics 'self level' and 'social-reference norm' were observed only once, we did not calculate correlations for these variables.

## **Acknowledgement**

We would like to thank Malte Klimczak for his support in planning and conducting the study and Lena Hondrich for her comments that greatly improved the manuscript.

## **Disclosure statement**

No potential conflict of interest was reported by the authors.

## **Funding**

This work was supported by the German Research Foundation [grant number KL 1057/10-3], [grant number BL 275/16-3], [grant number LE 2619/1-3].

## **Notes on contributors**

Petra Birgitta Pinger received her Master of Science degree in Psychology from the Maastricht University. Currently, she is a PhD candidate at the German Institute for International Educational Research (DIPF) in Frankfurt am Main. Her research is centred on the implementation and effectiveness of formative assessment.

Katrin Rakoczy is a postdoctoral scientist at the Department for Research on Educational Quality and Evaluation of the German Institute for International Educational Research (DIPF) in Frankfurt am Main. She received a diploma in Psychology from the Technical University of Dresden and did her PhD at the Goethe University in Frankfurt. She is primarily interested in research on instructional quality and the question how it influences student learning

Michael Besser is an assistant professor at the Faculty of Mathematics, Science and Technology at the University of Education in Freiburg, Germany. He has studied Mathematics and German for becoming a secondary teacher and has finished his PhD in Didactics of Mathematic at the University of Kassel. His research focuses on the impact of teachers' competence on the quality of teaching, on ways of implementing competency-oriented teaching of mathematics into school and on assessing learners' competencies.

Eckhard Klieme is Director of the Department for Research on Educational Quality and Evaluation at the German Institute for International Educational Research (DIPF). He is Full Professor of Educational Science at Goethe University and received a diploma in Mathematics and a doctorate in Psychology from the University of Bonn. His current research is focused on teaching quality and school effectiveness.

## References

- Ames, C. (1992). Classroom: Goals, structures, and student motivation. *Journal of Educational Psychology*, 84, 261–271.
- Bangert-Drowns, R. L., Kulik, C. C., Kulik, J. A., & Morgan, M. T. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research*, 61, 213–238. doi:10.3102/00346543061002213
- Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy & Practice*, 18, 5–25. doi:10.1080/0969594X.2010.513678
- 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). New York, NY: Springer.
- Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy, and Practice*, 5, 7–74. doi:10.1080/0969595980050102
- Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability*, 21, 5–31. doi:10.1007/s11092-008-9068-5
- 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, NY: Springer.
- Blum, W., Galbraith, P. L., Henn, H.-W., & Niss, M. (2002). ICMI study 14: Applications and modelling in mathematics education – Discussion document. *Educational Studies in Mathematics*, 51, 149–171.
- Cianci, A. M., Schaubroeck, J. M., & McGill, G. A. (2010). Achievement goals, feedback, and task performance. *Human Performance*, 23, 131–154. doi:10.1080/08959281003621687
- Cizek, G. J., Fitzgerald, S. M., & Rachor, R. E. (1996). Teachers' assessment practices: Preparation, isolation, and the kitchen sink. *Educational Assessment*, 3, 159–179. doi:10.1207/s15326977ea0302\_3
- 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 students' science understanding? *American Educational Research Journal*, 52, 1133–1159. doi:10.3102/0002831215596412
- 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
- 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 , 269–290. doi:10.1080/01443410.2013.785384

Hattie, J. (2009). Visible learning. A synthesis of over 800 meta-analyses relating to achievement. New York, NY: Routledge.

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. (2015). Implementing curriculum-embedded formative assessment in primary school science classrooms. *Assessment in Education, Principles, Policy & Practice*. Advance online publication. doi:10.1080/0969594X.2015.1049113

Huth, K. (2004). Entwicklung und Evaluation von fehlerspezifischem informativem tutoriellem Feedback (ITF) für die schriftliche Subtraktion [Development and Evaluation of bug-related tutoring feedback for written subtraction] (Doctoral dissertation). Retrieved from <http://www.qucosa.de/fileadmin/data/qucosa/documents/1243/1105354057406-4715.pdf>

Kingston, N., & Nash, B. (2011). Formative assessment: A meta-analysis and a call for research. *Educational Measurement: Issues and Practice*, 30, 28–37. doi:10.1111/j.1745-3992.2011.00220.x

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.

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

Köller, O. (2005). Formative assessment in classrooms: A review of the empirical German literature. In OECD (Ed.), *Formative assessment: Improving learning in secondary classrooms* (pp. 265–279). Paris: OECD.

Krampen, G. (1987). Differential effects of teacher comments. *Journal of Educational Psychology*, 79, 137–146. doi:10.1037/0022-0663.79.2.137

Krapp, A. (1999). Interest, motivation and learning: An educational-psychological perspective. *European Journal of Psychology of Education*, 14, 23–40.

Krapp, A. (2002). Structural and dynamic aspects of interest development: Theoretical considerations from an ontogenetic perspective. *Learning and Instruction*, 12, 383–409.

Krapp, A. (2005). Basic needs and the development of interest and intrinsic motivational orientations. *Learning and Instruction*, 15, 381–395.

Krapp, A., & Prenzel, M. (2011). Research on interest in science: Theories, methods, and findings. *International Journal of Science Education*, 33, 27–50. doi:10.1080/09500693.2010.518645

Kulhavy, R. W., White, M. T., Topp, B. W., Chan, A. L., & Adams, J. (1985). Feedback complexity



and corrective efficiency. *Contemporary Educational Psychology*, 10, 285–291. doi:10.1016/0361-476X(85)90025-6

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

Lee, H., Feldman, A., & Beatty, I. D. (2011). Factors that affect science and mathematics teachers' initial implementation of technology-enhanced formative assessment using a classroom response system. *Journal of Science Education and Technology*, 21, 523–539. doi:10.1007/s10956-011-9344-x

Leiss, D., Schukajlow, S., Blum, W., Messner, R., & Pekrun, R. (2010). The role of the situation model in mathematical modelling – Task analyses, student competencies, and teacher interventions. *Journal für Mathematik-Didaktik*, 31, 119–141.

Maaß, K. (2006). What are modelling competencies? *Zentralblatt für Didaktik der Mathematik (ZDM)*, 38(2), 113–142.

Maaß, K. (2010). Classification scheme for modelling tasks. *Journal für Mathematik-Didaktik*, 31, 285–311. doi:10.1007/s13138-010-0010-2

Meyer, W.-U. (1982). Indirect communications about perceived ability estimates. *Journal of Educational Psychology*, 74, 888–897.

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

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. doi:10.1016/j.learninstruc.2006.07.003

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.

O'Donnell, C. L. (2008). Defining, conceptualizing, and measuring fidelity of implementation and its relationship to outcomes in K-12 curriculum intervention research. *Review of Educational Research*, 78, 33–84. doi:10.3102/0034654307313793

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.'* Materialien zur Bildungsforschung [Documentation of the instruments of the Swiss-German video study 'Instructional quality, learning patterns, and mathematical understanding.' Materials on educational research] (Vol. 13, pp. 109–111). Frankfurt am Main: GPF.

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., Leiss, D., & Blum, W. (in press). Formative assessment in mathematics instruction: Theoretical considerations and empirical results of the Co<sup>2</sup>CA project. In D. Leutner, J. Fleischer, J. Grünkorn, & E. Klieme (Eds.), *Competence assessment in education: Research, models, and instruments*. Berlin: Springer.

Rheinberg, F. (2006). Bezugsnorm-Orientierung [Reference orientation]. In D. H. Rost (Ed.), *Handwörterbuch Pädagogische Psychologie [Handbook of educational psychology]* (3rd ed., pp. 55–62). Weinheim: Beltz.

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, 55, 68–78.

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

Shih, S., & Alexander, J. (2000). Interacting effects of goal setting and self- or other-referenced feedback on children's development of self-efficacy and cognitive skill within the Taiwanese classroom. *Journal of Educational Psychology*, 92, 536–543. doi:10.1037//0022-0663.92.3.536

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

Stigler, J. W., Gallimore, R., & Hiebert, J. (2000). Using video surveys to compare classrooms and teaching across cultures: Examples and lessons from. *Educational Psychologist*, 35, 87–100.

Tierney, R. D. (2006). Changing practices: Influences on classroom assessment. *Assessment in Education*, 13, 239–264. doi:10.1080/09695940601035387

William, 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, NY: Lawrence Erlbaum Associates.

Wu, M. L., Adams, R. J., & Wilson, M. R. (1998). *ConQuest: Generalised item response modelling software*. Camberwell: Australian Council for Educational Research.

## Appendix

Pretest item (first published in Besser, Blum, & Klimczak, 2013).

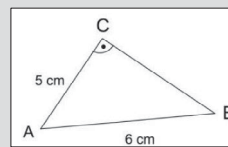
A broom is rested against a wall as shown below.  
Broom, wall and bottom form a triangle. Mark the triangle in the picture and give names to the sides.



Posttest item: technical task (first published in Besser, Blum, & Klimczak, 2013).

Calculate the length of the side a =

a = \_\_\_\_\_

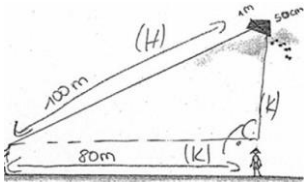


Posttest item: modelling problem (first published in Besser, Blum, & Klimczak, 2013).

On May 1<sup>st</sup> people in Bad Dinkelsdorf dance around a so called “Maibaum”. This is a tree which has a height of 8 m. While dancing, the people hold bands in their hands. These bands are 15 m long. How far away from the “Maibaum” are the people at the beginning of the dance?



Fictive example of the second diagnostic tool.

<div>Task 1</div> <p>Volker has been given a kite. The kite has a length of 1 m and a width of 50 cm. He flies the kite together with his friend Susanne. Both are placed 80 m from one another. The rope of the kite has a length of 100 m. Susanne is placed directly below the kite.</p> <p>What's the height of the kite at this moment?</p> <div><p>(not true to scale)</p></div> <div><math display="block">100^2 + 80^2 = x^2 \quad   \sqrt{\phantom{x}}</math><math display="block">x = \sqrt{100^2 + 80^2} \quad \checkmark</math><math display="block">x = 128 \text{ m} \quad \checkmark \rightarrow x = \sqrt{10000 + 6400}</math></div>	
<div><div>YOUR PERSONAL FEEDBACK</div><div>You are already quite good at dealing with the following topics:</div><div><div><i>I k-n.ow thca- mat'h, W not your b-ia- I thca-yow CU"e-o-w c;i, w!</i></div></div></div>	
<div><div>You can still improve at dealing with the following topics if concentrating on my hints:</div><div><i>Theorem of Pythagoras</i></div></div>	<div><div>Hints on how you can improve:</div><div><i>AMct your for help (for Le-our mat'h,w Tium/ JCM'W')</i></div></div>
<div>!! Please start working on your exercise now !!</div>	