Learning and Instruction 60 (2019) 154–165

Contents lists available at ScienceDirect

Learning and Instruction

journal homepage: www.elsevier.com/locate/learninstruc

Formative assessment in mathematics: Mediated by feedback's perceived usefulness and students' self-efficacy

Katrin Rakoczy[a,*], Petra Pinger[a], Jan Hochweber[b], Eckhard Klieme[a], Birgit Schütze[c], Michael Besser[d]

[a] German Institute for International Educational Research, Schloßstraße 29, D-60486 Frankfurt am Main, Germany
[b] The University of Teacher Education, Notkerstrasse 27, CH-9000 St. Gallen, Switzerland
[c] The University of Münster, Fliednerstr. 21, D-48149 Münster, Germany
[d] Leuphana University of Lüneburg, Scharnhorststraße 1, D-21335 Lüneburg, Germany

ARTICLE INFO

Keywords:
Formative assessment
Intervention study
Mediation
Perceived usefulness
Self-efficacy

ABSTRACT

Although formative assessment is regarded as a promising way to improve teaching and learning, there is considerable need for research on precisely *how* it influences student learning. In this study we developed and implemented a formative assessment intervention for mathematics instruction and investigated whether it had effects on students' interest and achievement directly and via students' perception of the usefulness of the feedback and their self-efficacy. We conducted a cluster randomized field trial with pretest and posttest. The 26 participating classes were randomly assigned to a control group or the intervention group. Results of path analyses indicate that feedback was perceived as more useful in the formative assessment condition, self-efficacy was greater, and interest tended to increase; learning progress did not differ between the groups. The assumed indirect effects were partly confirmed: formative assessment showed an indirect effect on interest via its perceived usefulness.

1. Introduction

In the Anglo-American literature formative assessment is considered a very promising way to improve teaching and learning (e.g., Black & Wiliam, 2009; Stiggins, 2006; Wiliam & Thompson, 2008). The detailed synthesis of 250 studies of formative assessment published by Paul Black and Dylan Wiliam (Black & Wiliam, 1998a, 1998b, 1998c) probably is the most frequently cited source to support this assumption. With their synthesis they made an indispensable contribution to research on formative assessment by structuring the heterogeneous area of research, initiating several attempts to develop more sophisticated and applicable definitions of formative assessment (e.g., Bennett, 2011; Dunn & Mulvenon, 2009; Wiliam & Thompson, 2008), and providing a conceptual basis for interventions to implement formative assessment in the classroom (e.g., Educational Testing Service, ETS, 2009; Wiliam, Lee, Harrison, & Black, 2004). Despite their enormous contribution to this area of research, the trustworthiness of the review as a source of empirical evidence of a strong effect of formative assessment on learning is debatable (Bennett, 2011; Dunn & Mulvenon, 2009; Kingston & Nash, 2011). Although Black and Wiliam (1998c) clearly stated that they did not apply any quantitative meta-analytic techniques to the data they gathered (see p.53), they reported in another paper (Black & Wiliam, 1998a) an effect size of classroom assessment on student achievement of between .4 and .7 standard deviations, which was cited over 2700 times between 1998 and 2011 (Kingston & Nash, 2011). The broad definition of formative assessment used in their synthesis covered a very heterogeneous body of research which did not reliably answer the question of whether formative assessment affects learning (Bennett, 2011; Black & Wiliam, 1998a; Kingston & Nash, 2011). According to critics, a more appropriate conclusion for Black and Wiliam to have drawn would have been that more empirical research in the area of formative assessment was needed (Dunn & Mulvenon, 2009). More precisely, such research should (1) put greater care into the evaluation of the sources of evidence and the attributions made about them, and (2) develop a clearer definition of what is meant by formative assessment (Bennett, 2011).

The first issue was addressed by Kingston and Nash (2011), who conducted a meta-analysis with five strict criteria for the inclusion of studies, which resulted in a study base of only 13 studies with 42 effect sizes. They found a much smaller but still meaningful mean effect size

* Corresponding author.
E-mail addresses: rakoczy@dipf.de (K. Rakoczy), pinger@dipf.de (P. Pinger), Jan.Hochweber@phsg.ch (J. Hochweber), klieme@dipf.de (E. Klieme), harksb@uni-muenster.de (B. Schütze), michael.besser@leuphana.de (M. Besser).

https://doi.org/10.1016/j.learninstruc.2018.01.004
Received 12 April 2017; Received in revised form 8 December 2017; Accepted 18 January 2018
Available online 15 February 2018

of .20 for formative assessment. Although their meta-analysis helped debunk the myth that formative assessment has an average effect size of .4–.7, several authors challenge the accuracy of the new baseline of .20 for the average efficacy of formative assessment interventions (e.g., Briggs, Ruiz-Primo, Furtak, Shepard, & Yin, 2012; McMillan, Venable, & Varier, 2013). They identified methodological problems that call the accuracy of this estimate into question and mitigate their conclusions. Briggs and colleagues focus on study retrieval, application of inclusion criteria, computation of effect sizes, and choice of outcome measures as potential methodological problems. McMillan and colleagues believed the most significant shortcomings of Kingston and Nash's meta-analysis were the lack of attention to the methodological quality of the studies (i.e., unit of analysis, selection, and instrumentation) and the lack of consideration of the specific nature of formative assessment under investigation in each study. The authors of both studies conclude that higher quality studies are needed but admit that this is not an easy task given the applied nature of the research.

The second issue involves specifying exactly what effective formative assessment constitutes and how the reported gains in student learning can be achieved (Wiliam & Thompson, 2008). To understand better how learning gains can be achieved through formative assessment, underlying mechanisms that are supposed to cause and explain the intended effects on students' interest and achievement need to be evaluated (Bennett, 2011).

In our study we addressed the second issue and contribute to a deeper understanding of formative assessment and its impact on learning gains by identifying central components of formative assessment and analyzing underlying mechanisms of their impact. Hence, we identified elicitation of diagnostic information and feedback as central components of formative assessment (see section 2). Referring to criteria to design effective elicitation and feedback (see section 3) we developed a formative assessment intervention in which teachers implemented both components in their instruction. We determined whether and how the formative assessment intervention affected students' interest and achievement by investigating students' perception of its usefulness and their self-efficacy as intervening variables (see section 4). We conducted the study in mathematics instruction because more than 80% of class time in mathematics is spent working on tasks and solving problems (Hiebert et al., 2003); therefore, feedback on performance in this subject becomes particularly important for the students.

2. Elicitation of diagnostic information and feedback as central components of formative assessment

According to Andrade (2010) formative assessment has the two main purposes of "(1) providing information about students' learning to teachers and administrators in order to guide them in designing instruction; and (2) providing feedback to students about their progress in order to help them determine how to close any gaps between their performance and the targeted learning goals" (p. 344f; a similar definition can be found in Stiggins, 2006). To make assumptions about how teachers and students should respond to information they receive, a theory of action would be helpful (Bennett, 2011). The framework of Wiliam and Thompson (2008) can be seen as a rudimentary theory of action for providing feedback (Bennett, 2011). It suggests that formative assessment can be conceptualized as consisting of five key strategies: 1. to clarify and share learning intentions and criteria for success in order to determine the direction in which learners are heading; 2. to elicit evidence of students' understanding (assessment) in order to determine the areas in which they are reaching their learning goals; 3. to provide feedback that pushes learners forward; 4. to encourage students to be instructional resources for one another; and 5. to motivate students to take responsibility for their learning (Wiliam & Thompson, 2008). The first three key strategies of formative assessment basically reflect the feedback questions of Hattie and Timperley (2007): "Where am I going?", "How am I going?", and "Where to next?", asking for the learning goal, learning progress, and learning strategies. Key strategies 4 and 5 illustrate that the questions of learning goal, learning progress, and learning strategies can be addressed not only by teachers but also by peers or the learners themselves (e.g., Clark, 2012; Nicol & Macfarlane-Dick, 2006). Peer-assessment and self-assessment have been found to have considerable positive effects on outcome variables (for peer-assessment, see e.g., Van Zundert, Sluijsmans, & van Merriënboer (2010); for self-assessment, see e.g., Panadero and Jonsson (2013), Panadero, Jonsson, & Botella (2017)). In the present study, however, we focus on teachers as actors to implement the first three formative assessment strategies.

Coming back to the question of what formative assessment constitutes, the particular importance of eliciting evidence and providing feedback has been emphasized by many authors (e.g., Black & Wiliam, 1998a, 2009; Hattie, 2003; Kingston & Nash, 2011; Sadler, 1998; Stiggins, 2006). According to McMillan et al. (2013) it is apparent that in most studies in the field emphasis has been on gathering data and providing feedback, and little emphasis has been on instructional correctives. Eliciting evidence and providing feedback are closely connected and not clearly separable. Basically, educational assessment involves not only creating opportunities to gather evidence, collecting it, and interpreting it but also acting on interpretations, which may include feedback (Bennett, 2011). However, feedback addressing the three aforementioned feedback questions of Hattie and Timperley (2007) should include diagnostic information on how the learner is progressing.

3. Design of effective formative assessment

The relevance of feedback on elicited information within formative assessment raises the questions as to how information should be elicited and, particularly, how feedback should be delivered so as to support student learning (Pellegrino, Chudowsky, & Glaser, 2001). In the following, we restrict our description to elicitation and feedback by the teacher.

3.1. Eliciting evidence of student learning by the teacher

The following criteria are considered important to obtain reliable and valid information on student learning: (a) Diagnostic instruments should be aligned with the instruction (Wilson & Sloane, 2008) and specific domain (Bennett, 2011) in which they are applied and with a sound cognitive domain model (Pellegrino et al., 2001); (b) These instruments should be embedded in an ongoing, interconnected series of assessments so that patterns in student learning can be identified (Heritage, 2007; Stiggins, 2006); (c) The reasonably deep understanding of students' cognition needs to be combined with general principles, strategies, and techniques to be able to deliver supportive feedback (see section 3.2) and to be effective (Bennett, 2011); and (d) teachers need help to implement formative assessment in their instructional practices (Bennett, 2011).

3.2. Teacher feedback on elicited information

Hattie and Timperley (2007) emphasize that the main purpose of feedback is to highlight the discrepancy between current understanding and performance on one hand and the learning goal on the other, and to encourage and enable students to reduce the discrepancy. Similarly, Shute (2008, p. 154) defines formative feedback as "information communicated to the learner that is intended to modify his or her thinking or behavior for the purpose of improving learning."

To enable students to reduce the discrepancy between the learning goal and their current performance on a mathematics test they should be informed about whether or not they applied the mathematical operations needed to solve the tasks on the test (strengths and weaknesses), and they should be provided with information that may be

useful for determining which strategies would be most appropriate for solving the tasks (see Harks, Rakoczy, Hattie, Besser, & Klieme, 2014; Rakoczy, Harks, Klieme, Blum, & Hochweber, 2013). By informing the learner about his or her strengths, weaknesses, and strategies, such feedback – referred to as process-oriented feedback in the following – helps the learner answer the three feedback questions of Hattie and Timperley (2007; Where am I going? How am I going? Where to next?; see section 2).

Feedback can be provided at the 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 the task level, process level, and self-regulatory level by referring to specific tasks and by focusing on cognitive and self-regulatory processes (Harks et al., 2014; Rakoczy et al., 2013).

Process-oriented feedback has been shown to support student learning in mathematics better than social-comparative feedback in an experimental setting (Harks et al., 2014; Rakoczy et al., 2013) and realizes two of Narciss' (2008) elaborated feedback components: 'knowledge about mistakes' by informing about weaknesses and 'knowledge about how to continue' by giving strategies. Thereby, it serves a corrective function according to Shute (2008). At the same time it fulfills the following basic motivational functions listed by Narciss (2008): provide an incentive (by rendering the result visible), facilitate task completion (by offering suggestions to overcome difficulties), enhance self-efficacy (by making it possible to master tasks), and contribute to mastery experience that can be attributed to personal causation. For a detailed description of the facets of process-oriented feedback and their theoretical foundation, see Harks et al. (2014) or Rakoczy et al. (2013).

4. How formative assessment influences student learning

According to Andrade (2010) the essence of formative assessment is informed action. That is, teachers must know how to respond to the information obtained through assessment and adjust their instruction according to students' needs; students must be equipped with strategies and have the motivation needed to improve their work and deepen their understanding after receiving feedback. In other words, formative assessment does not simply result in better learning, but rather, drawing upon the theory of action, formative assessment is assumed to initiate particular actions which, in turn, lead to better learning outcomes (Bennett, 2011). Concerning informed action of students – the focus of the present study – it cannot simply be assumed that students who are provided with feedback within a formative assessment intervention will know what to do with it or will automatically respond to it as intended (Sadler, 1998). Rather, feedback has a certain functional significance for the learner depending on his or her perception and interpretation of it (Black & Wiliam, 2012; Brookhart, 1997). Therefore, a theoretical model is needed of the mediations through which feedback influences cognition in general and the learning process in particular (Perrenoud, 1998; Rakoczy et al., 2013). According to Stiggins (2006), a productive way to respond to formative assessment involves students perceiving the feedback provided as useful ("I understand these results. I know what to do next to learn more."), and feeling a high level of self-efficacy regarding the forthcoming task ("I can handle this. I choose to keep trying."). In the following sections we describe in more detail why and how the perceived usefulness of feedback and students' self-efficacy are assumed to explain effects of formative assessment interventions on students' interest and achievement.

4.1. Perceived usefulness

Feedback helps students identify the particular aspects of their work that need attention (Sadler, 2011). It fulfills cognitive and motivational functions (Narciss, 2008; see section 3.2) by initiating cognitive and behavioral adaptive reactions for error correction, which are related to interest as well as future achievement (Tulis, Steuer, & Dresel, 2016). For this purpose, feedback needs to be accepted as useful for cognitive and behavioral adaptive reactions (Mouratidis, Vansteenkiste, & Lens, 2010). Conveniently, Panadero and Jonsson (2013) included students' reflection on the usefulness of feedback as a mediating factor in their model for the effectiveness of using rubrics as a self-assessment instrument. Empirical evidence for the mediating role of perceived usefulness was provided by Harks et al. (2014) and Rakoczy et al. (2013). Process-oriented feedback perceived as useful seemed to help students understand their mistakes and identify strategies to proceed (Narciss, 2008; Shute, 2008). It is also connected with students' perceived competence support (Rakoczy et al., 2013) and in the terminology of Carver and Scheier (1998) it can be referred to as a discrepancy-reducing feedback loop.

Facilitation of task completion by feedback perceived as useful is believed to be connected with the self-efficacy enhancing function of feedback (Narciss, 2008). In other words, the perceived usefulness of feedback probably is not only an important step for its actual use in error correction, but its anticipation should result in higher self-efficacy. That is, students who perceive feedback as useful for error correction should report greater self-efficacy regarding a forthcoming task or test because they see the possibility to improve after a mistake by actively analyzing and correcting it.

4.2. Self-efficacy

Self-efficacy is one's goal-referenced, relatively context-specific, and future-oriented beliefs about one's competence. These beliefs are malleable due to their dependence on tasks (Schunk & Pajares, 2009). Self-efficacy frequently has been shown to affect students' interest and achievement (Jiang, Song, Lee, & Bong, 2014; Pajares, 1996; Schunk, 1995). It can enhance interest and academic achievement by influencing students' effort, persistence, perseverance, and use of strategies (Pajares, 1996).

Helping students develop self-efficacy for mastering mathematics is one way teachers can enhance learning outcomes (Brookhart, 1997; Wigfield, Eccles, & Rodriguez, 1998). Even though the most important source of self-efficacy is one's prior performance (Bandura, 1997), social persuasion in the form of positive feedback also has an important impact on self-efficacy (Schunk, 1995). That is, feedback – the core component of formative assessment – has the potential to foster students' self-efficacy. Feedback is interpreted by individuals, and these interpretations provide the information upon which one judges one's competence (Pajares, 1996). As interpretation of feedback should cultivate people's beliefs in their capabilities and ensure that the envisioned success is attainable (Schunk & Pajares, 2009), self-efficacy enhancing feedback should focus not only on the errors students make, but also on their strengths and provide them with strategies to adapt their previous approach to reduce the discrepancy between their learning goal and their current performance (Hattie & Timperley, 2007; Schunk & Swartz, 1993; Shute, 2008).

Yin et al. (2008) investigated the impact of formative assessment on self-efficacy but could not find a positive effect. They explained the lack of empirical evidence with poor implementation of formative assessment (Furtak et al., 2008). Meta-analyses of the impact of self-assessment, however, revealed an effect size of 0.73 for self-assessment interventions on self-efficacy (Panadero et al., 2017). Concerning rubrics as a self-assessment instrument, however, students need to be confronted with teachers' feedback regarding their performance (Panadero & Jonsson, 2013).

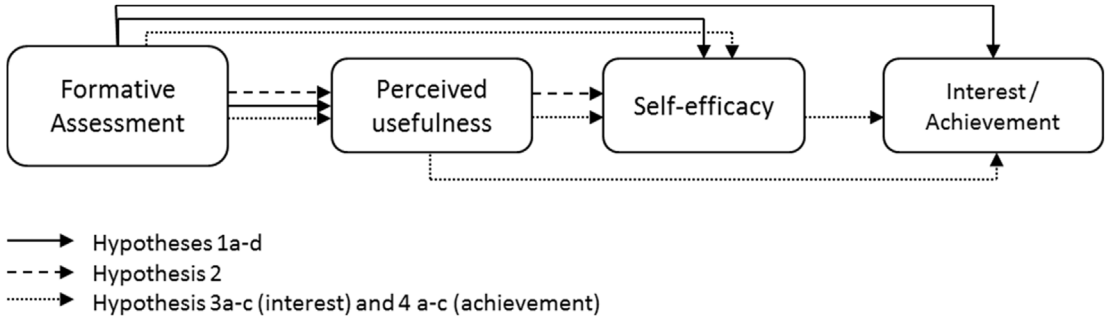

Fig. 1. Overview of hypotheses.

5. Research questions and hypotheses

In the present study we investigated whether and how a formative assessment intervention in mathematics classes affected students' interest and achievement and compared results to those of a control group in which no formative assessment practices were conducted. Drawing on the theoretical insights and empirical evidence outlined in the previous sections, this study was guided by the following research questions and hypotheses (for an overview, see Fig. 1).

1. Does formative assessment have an impact on students' perception of the usefulness of the feedback, their self-efficacy, interest, and achievement?

Hypotheses. We expect formative assessment to have an effect on students' perception of the usefulness of the feedback, their self-efficacy, interest, and achievement (Hypotheses 1a-d).

2. Does formative assessment have an indirect impact on students' self-efficacy via their perception of the usefulness of the feedback?

Hypothesis. We expect formative assessment to have an indirect effect on students' self-efficacy via its perceived usefulness. That is, students in the formative assessment condition should perceive their feedback as more useful than students in the control condition, and as a result they should become more confident with regard to their future achievement in mathematics.

3. Does formative assessment have an impact on students' interest via their perception of the usefulness of the feedback and their self-efficacy?

Hypotheses. We expect formative assessment to have an indirect effect on students' interest via the students' perception of the usefulness of the feedback and their self-efficacy. That is, in addition to the path described in Hypothesis 2, a higher level of self-efficacy is assumed to be related to growth in individual interest, leading to a positive indirect effect on interest (3a). The impact of perceived usefulness of the feedback on interest is explained not only by its self-efficacy enhancing function but also by other motivational processes such as supporting the need for competence. Therefore, we expect formative assessment alone to have an additional indirect effect on interest via its perceived usefulness (3b). Finally, we expect formative assessment to have an indirect effect on interest via self-efficacy (and not via its perceived usefulness) as the information provided in the feedback is expected to influence students' beliefs about their competence, which are connected to interest (3c).

4. Does formative assessment have an impact on students' achievement via students' perception of the usefulness of the feedback and their self-efficacy?

Hypotheses. We expect formative assessment to have an indirect effect on students' achievement via their perception of the usefulness of the feedback and their self-efficacy. That is, in addition to the path described in Hypothesis 2, greater self-efficacy is assumed to be related to greater individual achievement, leading to a positive indirect effect on achievement (4a). Analogously to Hypothesis 3b, we expect an additional indirect effect only via students' perception of the usefulness of the formative assessment (4b). Finally, analogously to Hypothesis 4c we assume an additional indirect effect via self-efficacy (and not via perceived usefulness of feedback) on achievement (4c).

6. Methods

This study is based on data from the project "Conditions and Consequences of Classroom Assessment (Co^{2}CA)" which was conducted by the German Institute for International Educational Research, the University of Kassel, and the Leuphana University of Lüneburg.[1] Our study is part of a larger study in which the impact of formative assessment on student learning is investigated. The design of the complete project is reported in Rakoczy, Klieme, Leiß, and Blum (2017). The methods described here are limited to the part of the design of the complete study used in our analyses.

6.1. Participants

In the study, 26 teachers (69% female) from 18 middle track schools (Realschulen) in Hesse, Germany took part with their 9th grade mathematics classes.[2] We took into account that we cannot prevent teachers from the same school from discussing their training; therefore, we assigned classes from the same school to the same experimental group so as to avoid diffusion within schools. Of the 620 students 45% were female and the mean age at the beginning of the study was 15.1 years (SD = 7.46 months). Participation in the study was voluntary.

6.2. Design

We conducted a cluster randomized field trial with pretests and posttests during the 2010/2011 school year. Classes were randomly assigned to the intervention group (n = 11 classes, i.e. 259 students) or the control group (n = 15 classes, i.e. 361 students). In the intervention group, teachers administered a formative assessment intervention based on written process-oriented feedback in mathematics; in the control group, no specific performance assessment was conducted or feedback given. In both groups the mathematical content and tasks the students worked on were standardized. Conditions were realized by teachers according to the training they received (the content of the teacher training is described in section 6.4).

[1] The project was supported by grants from the German Research Foundation (DFG, KL 1057/10, BL 275/16 and LE2619/1).

[2] The sample originally consisted of 29 teachers with their classes, but three classes had to be excluded due to a lack of implementation veracity (see Pinger et al., 2016).

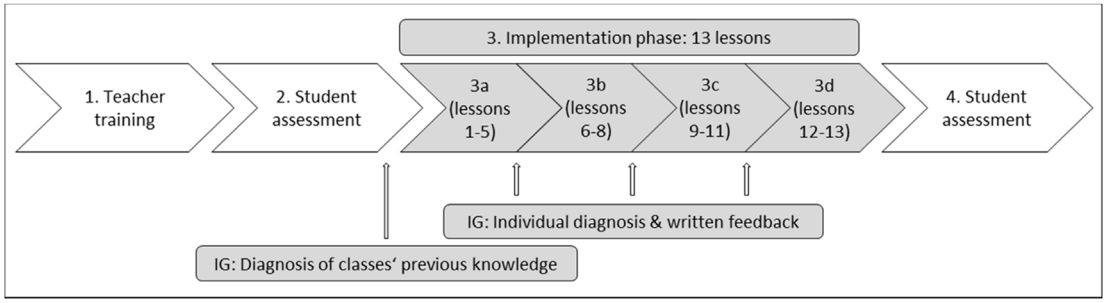

Fig. 2. Procedure.

6.3. Procedure

The cluster randomized field trial covered the first 13 lessons of the teaching unit "Theorem of Pythagoras". The 13 lessons took place over approximately three weeks and consisted of four phases (see Fig. 2). In phase 1, teachers participated in the teacher training session of the condition to which they were assigned (see section 6.4). In phase 2, students were assessed in four steps in the lesson before the teaching unit started: first, students were informed about the content of a forthcoming mathematics test. Second, their self-efficacy regarding the forthcoming test and their interest were assessed using a questionnaire. Third, their achievement in mathematics was measured in a pretest. Fourth, students' perception of the usefulness of their teacher's feedback as usually given to their class was recorded on a questionnaire and examined. In phase 3, the implementation phase, all the teachers administered the mathematical tasks and problems they were given during their training. The implementation phase was divided into four segments representing learning progression according to the cognitive domain model used (see section 6.4.1): a) introduction, proof, technical tasks, b) dressed-up word problems, c) modeling problems, and d) consolidation. In the intervention group, before starting the teaching unit the teachers received an overview of their students' prior knowledge of Pythagoras as assessed in the pretest. The teachers assessed students' performance at the end of each phase at three predefined points in time (in the 5th, 8th, and 11th lessons) and provided students with written process-oriented feedback in the following lesson using the diagnostic and feedback tool developed according to the design principles described in section 3 (see section 6.4.2). In phase 4, students were assessed again in four steps: first, they were given information on the content of a forthcoming posttest. Second, their self-efficacy and interest in the posttest were assessed using a questionnaire. Third, a posttest measuring students' achievement in mathematics was administered. Fourth, the students' perception of the usefulness of the teacher's feedback was recorded on a questionnaire and examined.

6.4. Teacher training and experimental conditions

To ensure that the subject-specific content and mathematical tasks and problems during the 13 lessons of the study were comparable among participating classes, all the teachers took part in one half-day training session on content and mathematical tasks and problems (see section 6.4.1). Teachers of the intervention group participated in another half-day teacher training session on formative assessment (see section 6.4.2).

6.4.1. Teacher training on content and mathematical tasks and problems

All the teachers were introduced to the subject-specific content and were provided with the mathematical tasks for the first 13 lessons of the teaching unit on Pythagoras' theorem. Teachers also were trained to use a didactic approach focusing on students' ability to apply mathematical tools to real-world problems ("mathematical modeling"; see Blum & Leiss, 2007; Leiss, Schukajlow, Blum, Messner, & Pekrun, 2010). Based on this approach as a cognitive domain model and its description of learning progression, the 13 lessons of the study were divided into four phases (see 3a-3d in section 6.3). Tasks in the pretest and posttest as well as on the diagnostic tool for the intervention groups were developed to assess students' achievement at the various stages of the learning progression.

6.4.2. Teacher training on formative assessment

In the formative assessment condition, teachers received training on the content of the intervention as well as on assessment of students' performance at the end of phases (a), (b), and (c), and they were instructed to give written process-oriented feedback. To this end, teachers were provided with a diagnostic and feedback tool which they were instructed to employ according to a partly standardized procedure. The application of the diagnostic and feedback tool at the end of each phase provided teachers with information on their students' performance level and students with individual profiles of strengths and weaknesses and ways to improve their performance.

The diagnostic and feedback tool consisted of an assessment part on the left and a feedback part on the right (see Fig. 3). The assessment part contained one or two mathematical tasks assessing knowledge of the content taught in the previous phase (two technical tasks at the end of phase (a), one dressed up word problem at the end of phase (b), one modeling problem at the end of phase (c)). At the end of the 5th, 8th, and 11th lessons teachers asked the students to complete within 15 min the tasks on the diagnostic and feedback tool. After the lessons the teachers assessed students' solutions and wrote in the feedback part of the tool individualized process-oriented feedback for students on their strengths ("In the following area(s) you are doing well"), weaknesses ("In the following area(s) you could improve"), and recommended strategies to continue ("This is how you can improve"). To support teachers and to ensure quality we partly standardized the procedure. We developed a list of cognitive processes and operations based on cognitive task analyses which were needed to solve the respective diagnostic task (e.g., identify catheti and hypotenuse in a right-angled triangle). Each process and operation could be fed back as a strength or weakness depending on whether or not the student had mastered it. For each process and operation a strategy or hint as to how to continue was provided if the respective process had not been mastered. To keep the feedback concise and understandable teachers were asked to summarize strengths across sub-competences and to choose the weaknesses they believed to be most significant. At the end of the feedback part students were asked to complete an additional similar task and apply the strategies provided in the feedback.

To summarize, our diagnostic and feedback instrument met the requirements formulated in section 3.1. It contained diagnostic tasks aligned with the modeling circle as a cognitive domain model in mathematics instruction (a) and could be applied at multiple points in time to describe learning progression (b). The design of the process-oriented feedback was based on recommendations found in the literature on feedback (c), and the teachers were trained to implement the diagnostic and feedback tool (d).

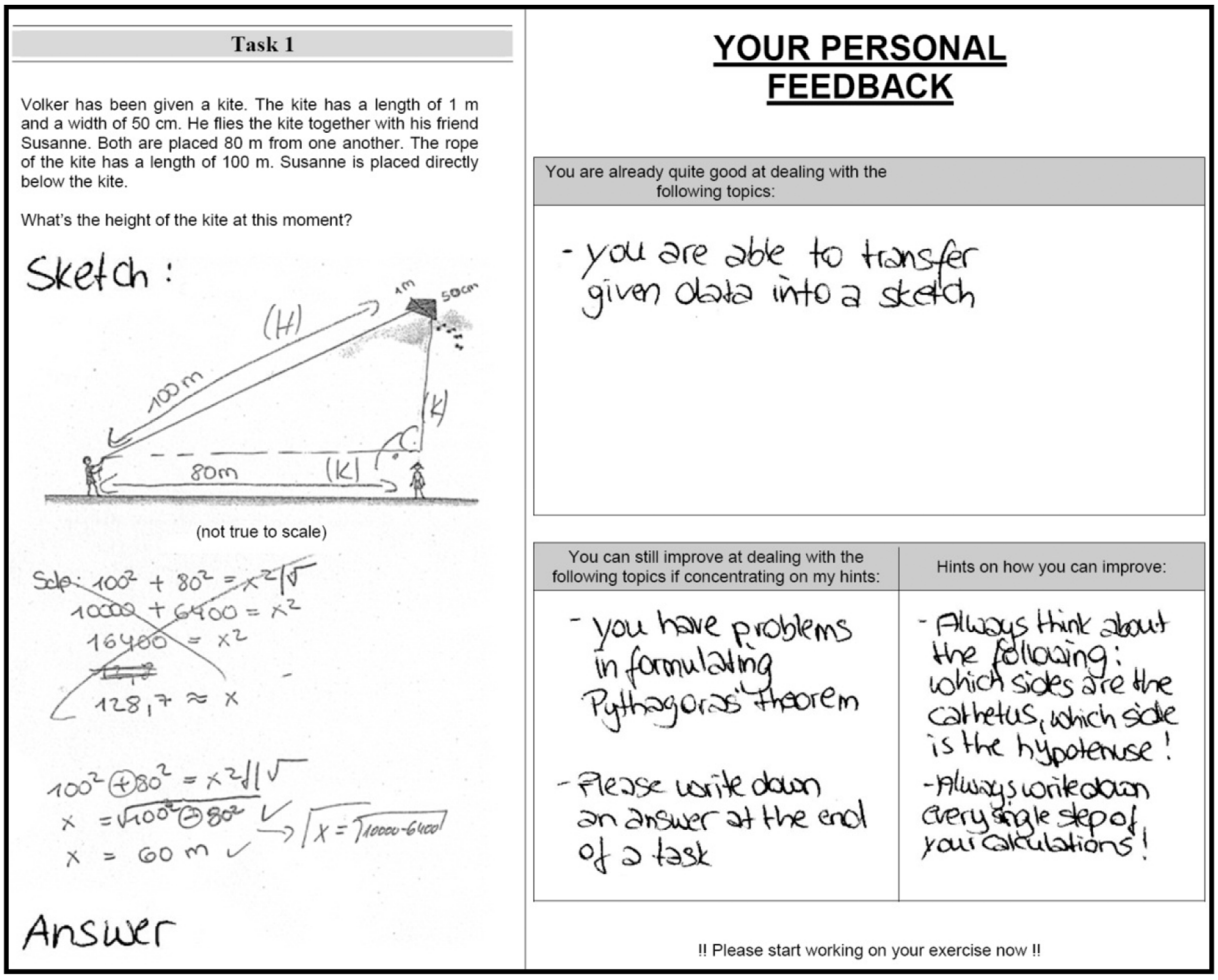

Fig. 3. Example of the diagnostic and feedback tool (used at the end of phase (b)).

6.5. Measures

6.5.1. Perceived usefulness

Students' perception of the usefulness of the formative assessment was self-reported on a five-item scale adapted from Dresel and Ziegler (2002). Students indicated on a four-point scale ranging from 0 (completely disagree) to 3 (completely agree) the extent to which feedback in the classroom helped them learn where and how they could improve. The key item was "The feedback helps me recognize where I can improve" (for all items see Appendix A). Internal consistency of the scale was Cronbach's α .86 in the pre-questionnaire and .85 in the post-questionnaire.

6.5.2. Self-efficacy

Students' self-efficacy regarding the forthcoming test was self-reported on a four-item scale ranging from 0 (completely disagree) to 3 (completely agree). Students indicated the extent to which they believed they would succeed on the forthcoming test. The key item was "I believe that I will do well on this test" (for all items see Appendix A). Internal consistency of the scale was Cronbach's α .84 in the pre-questionnaire and .88 in the post-questionnaire.

6.5.3. Interest

To assess students' interest students rated on a four-point scale ranging from 0 (completely disagree) to 3 (completely agree) on a pre-intervention questionnaire and again on a post-intervention questionnaire how interesting they found the topic of the forthcoming test. The key item was "I like the topic of the test" (for all items see Appendix A). Internal consistency of the scale was Cronbach's α .83 in the pre-intervention questionnaire and .89 in the post-intervention questionnaire.

6.5.4. Achievement

Achievement in mathematics was assessed with 19 pretest and 17 posttest items. Test items consisted of technical and modeling items on the topic of Pythagoras' theorem (for examples, see Besser, Blum, & Klimczak, 2012). Items had been analyzed previously on the basis of a scaling sample (N = 1570) in the context of a preceding calibration study. The aim of this study was to develop and calibrate the tasks needed for our intervention study. The psychometric quality of the tasks was supported by applying item response theory to the data and content-specific competence models were developed. The tasks used in the intervention study were fixed on the competence dimension of the scaling study (for further information on the scaling study, see e.g., Bürgermeister, Klieme, Rakoczy, Harks, & Blum, 2014; Harks, Klieme, Hartig, & Leiß, 2014). A one-dimensional Rasch model was applied to the experimental data and parameters (i.e., achievement scores) of weighted likelihood estimators (WLE) were estimated. Analyses were conducted in ConQuest (Wu, Adams, & Wilson, 1998). Estimated reliability (EAP/PV) was .66 for the pretest and .74 for the posttest.

6.6. Data analyses

We chose a half-longitudinal design for our study (Cole & Maxwell, 2003; Preacher, 2015) to determine whether the formative assessment intervention influenced the change in students' self-efficacy via their

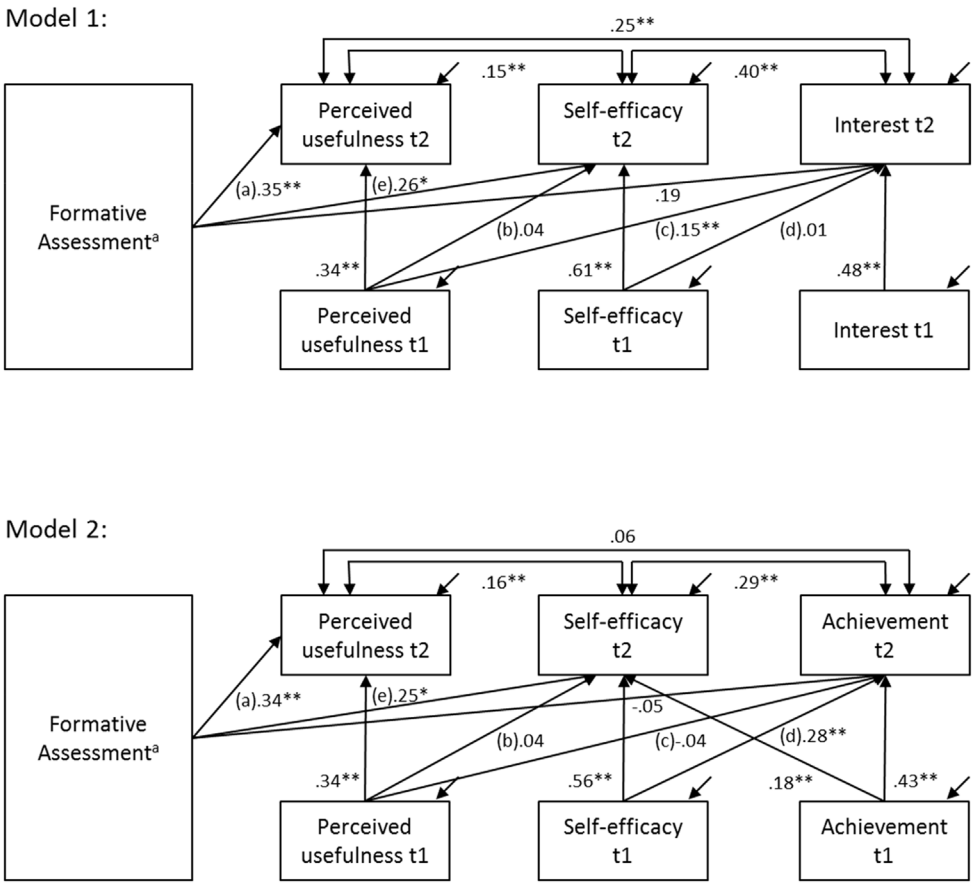

Fig. 4. Path models.

Note.

[a] Dummy-coded (intervention group= 1, control group= 0).

*p<.05, two-tailed, ** p < .01, two-tailed.

t1 = pre-questionnaire/test, t2 = post-questionnaire/test.

perception of the usefulness of the feedback and whether it influenced students' interest and achievement via their perception of the usefulness of the feedback and their self-efficacy. We computed direct, indirect, and total effects at the student level. With a half-longitudinal design an indirect effect can be estimated with data collected at two measurement points only: the effect of X_{t-1} on M_t is multiplied by the effect of M_{t-1} on Y_t to yield the indirect effect (Preacher, 2015). All analyses were conducted with MPlus 7.4 (Muthén & Muthén, 1998–2012).

To address our research questions we fitted two path models to the data – one for achievement and one for interest as outcomes (see Fig. 4).[3] In both models, the experimental group was entered as a dummy-coded predictor variable (0 = control group, 1 = intervention group). Perceived usefulness, self-efficacy, interest, and achievement at both measurement points were z-standardized based on their individual level mean and standard deviation in the pre-questionnaire/test. They were entered as manifest intervening variables or criteria. We estimated indirect effects according to the half-longitudinal design by multiplying (a) the effect of formative assessment on the intervening variable 1 in the post-questionnaire by (b) the effect of the intervening variable 1 in the pre-questionnaire on the intervening variable 2/outcome variable in the post-questionnaire/test, controlling for the intervening variable 2/outcome variable in the pre-questionnaire/test. For instance, the indirect effect of the formative assessment on self-efficacy at t2 was estimated by multiplying (a) the effect of formative assessment on perceived usefulness in the post-questionnaire, controlling for perceived usefulness of the feedback in the pre-questionnaire, by (b) the effect of perceived usefulness of the feedback in the pre-questionnaire on self-efficacy in the post-questionnaire, controlling for self-efficacy in the pre-questionnaire.[4] In the results section we describe which paths were multiplied to estimate each indirect effect.

As the variables under investigation had missing data in

[3] As the resulting path model would become rather complicated because of the large number of variables, we estimated two separate models for interest and achievement as outcome variables. However, a common model with both outcome variables confirmed the results of the separate models depicted here. The correlation between interest and achievement at t2 was significant but relatively weak (r = .133, SE = 0.061, p = .028).

[4] Unlike in the generic models discussed by Cole and Maxwell (2003) and Preacher (2015), we additionally controlled for the effects of formative assessment and any previous measurements of intervening variables up to the intervening variable/outcome of interest. For example, when estimating the indirect effect of formative assessment on interest at t2, we additionally controlled for (a) formative assessment when estimating the effect of its perceived usefulness at t1 on self-efficacy at t2 and (b) formative assessment as well as its perceived usefulness at t1 when estimating the effect of self-efficacy at t1 on interest at t2 (cf. Fig. 4).

9.8%–11.8% of the cases, we applied multiple imputation (MI) to replace each missing value with a set of predicted values (Schafer & Graham, 2002). We used the MI facility in Mplus, applying the standard unrestricted (H1) model of the relationships among all the variables (Asparouhov & Muthén, 2010). In addition to the variables under investigation, we included six individual-related background variables (e.g., gender) as auxiliary variables to obtain more precise estimates of the imputed variables and increase power (Collins, Schafer, & Kam, 2001). Because recent research has shown that relatively few imputations (e.g., five) may lead to serious power falloff (Graham, Olchowski, & Gilreath, 2007), we generated and analyzed 100 imputations for each variable.

Although teacher training on the provision of feedback was conducted at the teacher level, written feedback was given to students individually. Additionally, the intraclass correlations (ICCs) of both intervening and outcome variables indicated that the large majority of variation in the outcome variables occurred at the individual level (interest t2: ICC = .047; achievement t2: ICC = .117). Thus, we decided to focus on the theoretically well founded indirect effects at the individual level and not to analyze relationships separately at the teacher level (i.e., to conduct multilevel analyses). However, the clustering of the data was taken into account when computing standard errors and tests of model fit (by using the "type is complex" option in MPlus).[5]

7. Results

Before reporting on results of the path models to answer our research questions, we provide an overview in Table 1 of correlations, means, and standard deviations of the scales of the study.

The two path models used to address our four research questions (see Fig. 4) each show a satisfying fit to the data (Model 1: $\chi 2/df = 1.546$ (4.638/3); CFI = .996; RMSEA = 0.026; SRMR = 0.019; Model 2: $\chi 2/df = 1.215$ (2.429/2); CFI = .998; RMSEA = 0.017; SRMR = 0.008).

Concerning our first research question, the direct effects presented in Fig. 4 indicate that students indeed perceived feedback in the formative assessment condition as more useful, ($\beta = 0.348$, SE = 0.120, p = .004, Model 1, $\beta = 0.335$, SE = 0.083, p < .001, Model 2, Hypothesis 1a), [6] reported greater self-efficacy ($\beta = 0.261$, SE = 0.109, p = .016, Model 1, $\beta = .254$, SE = 0.116, p = .029, Model 2, Hypothesis 1b), and showed a marginally larger change in interest ($\beta = 0.193$, SE = 0.103, p = .062, Model 1, Hypothesis 1c). However, students did not differ with regard to their learning progress between both conditions ($\beta = -0.046$, SE = 0.170, p = .788, Model 2, Hypothesis 1d).[7]

Concerning our second research question, results presented in Fig. 4 indicate that students perceived feedback in the formative assessment condition as more useful (see Hypothesis 1a); however, unexpectedly, experiencing usefulness was not associated with greater self-efficacy ($\beta = 0.035$, SE = 0.042, p = .405, path b in Model 1, $\beta = .042$, SE = 041, p = .306, path b in Model 2). So, self-efficacy was not significantly indirectly fostered through formative assessment via its perceived usefulness ($\beta = .012$, SE = 0.014, p = .407, path a * path b in Model 1, $\beta = .014$, SE = .014, p = .311, path a * path b in Model 2, Hypothesis 2).

Concerning our third research question, the results of Model 1 (shown in Fig. 4) indicate that students' self-efficacy unexpectedly was not related to the change in their interest ($\beta = 0.007$, SE = 0.042, p = .862, path d). Accordingly, formative assessment had no indirect effect on students' interest via their perception of the usefulness of the feedback and self-efficacy ($\beta < 0.001$, SE = 0.001, p = .829, path a * path b * path d, Hypothesis 3a). Perceived usefulness of the feedback, however, was related to the change in students' interest as expected ($\beta = 0.146$, SE = 0.036, p < .001, path c). Hence, the path of formative assessment for interest change via its perceived usefulness (and not via self-efficacy) was statistically significant ($\beta = 0.051$, SE = 0.021, p = .017, path a * path c, Hypothesis 3b). As self-efficacy was not related to change in interest, the indirect effect of formative assessment via self-efficacy (and not via perceived usefulness of the feedback) was not statistically significant ($\beta = 0.002$, SE = 0.011, p = .868, path e * path d, Hypothesis 3c). Finally, the results of Model 1 indicate that the formative assessment intervention had a statistically significant total effect on change in interest ($\beta = 0.517$, SE = 0.203, p = .011).

Regarding the fourth research question, the results of Model 2 (shown in Fig. 4) indicate that, as expected, self-efficacy was related to students' achievement ($\beta = 0.280$, SE = 0.055, p < .001, path d). However, no indirect effect on students' achievement via their perception of the usefulness of the feedback and self-efficacy was found ($\beta = 0.004$, SE = 0.004, p = .325, path a * path b * path d, Hypothesis 4a). Perceived usefulness of the feedback was unexpectedly not related to students' achievement ($\beta = -0.042$, SE = 0.044, p = .349, path c). Therefore, the path of formative assessment on achievement via individuals' perception of the usefulness of the feedback (and not via their self-efficacy) also was not statistically significant ($\beta = -0.014$, SE = 0.016, p = .370, path a * path c, Hypothesis 4b). The indirect effect of formative assessment via self-efficacy (and not via perceived usefulness), however, almost reached statistical significance ($\beta = 0.071$, SE = 0.038, p = .063, path e * path d, Hypothesis 4c). Finally, the results of Model 2 indicate that formative assessment had no statistically significant total effect on achievement ($\beta = 0.212$, SE = 0.252, p = .401).

8. Discussion

We investigated the impact of a formative assessment intervention on interest and achievement directly as well as via students' perception of the usefulness of the feedback and their self-efficacy in half-longitudinal path analyses. We examined common and separate indirect effects via both intervening variables to determine which individual processes were activated by the formative assessment and to explain its impact on the outcome variables. Concerning direct effects our results indicate that students perceived teachers' feedback as more useful when teachers had employed the diagnostic and feedback tool according to how they had been instructed to use it during our teacher training in the formative assessment condition. Moreover, students reported being more confident regarding their achievement on a forthcoming test and tended to show greater interest in the topic of the test. Achievement, however, did not differ between the two conditions (Hypotheses 1a-d).

Students' perceptions of the usefulness of their teachers' feedback, unexpectedly, were not related to their self-efficacy, when the initial level of self-efficacy was controlled for. That is, the feedback provided was considered valuable and led to students' greater confidence in their performance on upcoming tasks; however, students' perception of the usefulness of the feedback and how they judged their competence were not connected. Consequently, no indirect impact on self-efficacy could be found (Hypothesis 2).

Contrary to Hypothesis 3a, no indirect effect via students' perception of the usefulness of the feedback and their self-efficacy was found

[5] When analyzing indirect effects, it would seem preferable to base the significance tests on a bootstrapping procedure rather than to rely on distributional assumptions that might be met only in large samples (e.g., MacKinnon, Lockwood, & Williams, 2004). Unfortunately, MPlus currently does not provide bootstrapping for hierarchical datasets (e.g., type = complex), and thus, we resort to standard errors and significance tests for the indirect effects based on the multivariate Delta method (e.g., Olkin & Finn, 1995).

[6] As prior achievement is the main source for the development of self-efficacy, in model 2 we included a path of achievement before the intervention on self-efficacy after the intervention. Therefore, the coefficients for the same paths differ slightly between the two models.

[7] We also tested the direct effects in separate regression models for each variable (controlling for previous measurements in correspondence with Fig. 4), and received almost exactly the same coefficients and standard errors.

Table 1
Correlations, means, and standard deviations of all variables in the study.

Variable	total sample										intervention group		control group	
	2	3	4	5	6	7	8	9	*M*	*SD*	*M*	*SD*	*M*	*SD*
1. Intervention group	-.00	-.01	.00	.01	.05**	.03*	.03*	-.01	–	–	–	–	–	–
2. Usefulness (pre)		.02	.06**	-.01	.12**	.02	.08**	-.02	3.09	.59	3.09	.57	3.09	.61
3. Self-efficacy (pre)			.08**	.16**	.01	.20**	.05*	.22**	2.36	.57	2.35	.60	2.37	.55
4. Interest (pre)				.07*	.05*	.07**	.02**	.11**	2.37	.63	2.37	.65	2.37	.62
5. Achievement (pre)					-.02	.18**	.04	.43**	−1.04	.92	−1.02	.84	−1.05	.96
6. Usefulness (post)						.07**	.14**	.02	2.86	.66	2.98	.63	2.77	.67
7. Self-efficacy (post)							.19**	.31**	2.74	.64	2.82	.66	2.68	.62
8. Interest (post)								.13**	2.63	.72	2.70	.74	2.58	.70
9. Achievement (post)									-.11	1.16	−0.13	1.17	−0.10	1.14

Note. Means and standard deviations are given for the total sample and separately for both experimental groups.
*p < .05, two-tailed, **p < .01, two-tailed.

on students' change in interest as self-efficacy in the pre-questionnaire was unexpectedly not related to interest. For the same reason, an indirect effect via self-efficacy alone is not statistically significant (Hypothesis 3c). Formative assessment impacted interest solely via its perceived usefulness (Hypothesis 3b) and the direct and indirect effects together resulted in a statistically significant total effect of formative assessment on interest.

We did not find an indirect effect of formative assessment on achievement via the two intervening variables (Hypothesis 4a), or via its perceived usefulness alone (Hypothesis 4b), as perceived usefulness of feedback unexpectedly was not related to achievement. Also, the indirect effect via self-efficacy barely failed significance (Hypothesis 4c). Here, the direct and indirect effects together did not result in a statistically significant total effect of formative assessment on achievement.

Interestingly, students' perception of the usefulness of the feedback was not related to their self-efficacy or achievement in the present study although the relationship between perceived usefulness of feedback and achievement had been found previously in an experimental setting (Rakoczy et al., 2013). It could be that the students found the strategies provided in the feedback to be useful but did not apply them. Perhaps they knew they would not improve in time for the forthcoming test and, therefore, their reported self-efficacy and achievement did not change. However, students who found the feedback helpful became more interested in the topic of the test. The feedback might have helped them appreciate new aspects of the topic and understand that their competence was malleable and thereby lead to greater interest.

In summary, our results indicate that the formative assessment intervention has the potential to change students' individual processes surrounding their perception of the usefulness of the assessment by eliciting diagnostic information and providing process-oriented feedback at several points in time during the teaching unit. This finding is in line with experience with process-oriented feedback in a laboratory setting (Harks et al., 2014; Rakoczy et al., 2013). Moreover, the formative assessment intervention helped students evaluate their own competence more positively. Therefore, process-oriented feedback as realized in the present study can be interpreted as a kind of social persuasion that fosters self-efficacy according to Schunk (1995).

The direct effects on interest and particularly on achievement were weaker and teachers need to consider the mediating effects of students' perceptions of the usefulness of their feedback and students' self-efficacy. They should create a climate of usability by emphasizing the importance of the feedback in general, contributing to an understanding of mistakes as opportunities to learn (Tulis, 2013), and helping students recognize the opportunities feedback bears for them to improve. Self-efficacy is known to be an important link between the learning context and learning outcome variables (Jiang et al., 2014) and the relationship frequently is explained by increased effort, persistence, and use of strategies. For educational practice it is important that teachers provide students with the opportunity to transform their individual self-efficacy into increased effort in the form of additional learning time or use of strategies. Taking the reported indirect effects together it becomes obvious that both outcome variables (interest and achievement) are fostered by different indirect paths: while for change in interest it seems particularly important to emphasize the use of strategies provided, for learning progress, it is particularly important to strengthen students' feeling of being competent and to provide them with additional learning opportunities or strategies in order for them to transform self-efficacy into increased effort.

Facing the range of effect sizes of formative assessment on learning reported in the introduction section the lack of a significant total effect on achievement in our study has to be considered further. Two points must be reflected: First, we compared the effects of process-oriented feedback given to students in an ecologically valid setting within the formative assessment intervention to the effects of feedback given naturally during instruction whereas in many previous studies (e.g., Gielen, Peeters, Dochy, Onghena, & Struyven, 2010; Strijbos, Narciss, & Dünnebier, 2010), the effects of feedback given to an experimental group was compared to the effects of not receiving any feedback at all. Second, methodology-wise, the absence of a total effect might emerge when two intervening variables show indirect effects that cancel each other out (Preacher & Hayes, 2008). This might be true for other intervening variables which were not observed in our study. Referring to the framework provided by Baron and Kenny (1986) and recent publications that have expanded on it, Fast et al. (2010) also reported indirect effects of a caring climate during instruction on performance in mathematics although they did not observe a total effect of caring on performance.

Concerning the interpretation of our results it is important to keep in mind that it was not possible in the present analyses to differentiate how teachers implemented the formative assessment intervention and how the quality of their implementation impacted students' achievement and interest in mathematics. In another publication on the same intervention study (Pinger, Rakoczy, Besser, & Klieme, 2016) we addressed this research question and found that students reached a higher level of achievement in mathematics and showed more interest when feedback was embedded in instruction and teachers encouraged students to use feedback.

8.1. Methodological limitations

When interpreting the results, some methodological limitations of the study need to be considered. First and most important, the sample size of 26 classes is rather small. Due to the small sample size, it would not have been possible to realize the half-longitudinal analyses in a multilevel path model (in addition to considerations described in section 6.6), let alone to apply doubly latent multilevel models (e.g., Morin, Marsh, Nagengast, & Scalas, 2014) to control for sampling and

measurement errors (see also Televantou et al., 2015). Instead, we focused on the individual level and took the nested data structure into account by correcting standard errors and tests of model fit. Certainly, however, the limited number of teachers involved in the study may have restricted the statistical power available. Calculations made in the CRTSize R package (Rotondi, 2015) indicated that the sample size was sufficient to obtain a power of 0.8 and to detect a moderate direct effect[8] of a binary predictor on an outcome in a cluster-randomized trial (given a group size of 20 and ICC = 0.10). Of course, calculating power accurately for indirect effects in such a setting would be much more complex.

Concerning the procedure, one limitation is that the teacher training was rather short. It would be interesting to investigate whether the intervention would have greater impact on student learning if teachers had more extensive training in assessing performance by administering diagnostic tasks and giving process-oriented feedback (a respective study was conducted by Besser & Leiss, 2014). In future research it would be desirable to include a greater number of classes with their teachers being provided with more extensive training.

A third limitation is that the measures we took could be improved. The intervening variables were identified in students' responses on a questionnaire before and after implementation of the teaching unit. Change was modeled by introducing the variables assessed before implementation as covariates. For future research it would be desirable for the variables of the mediational chain to be assessed not only in chronological order but with some period of time between assessments. Alternatively, future studies should examine students' underlying processes more thoroughly by assessing their needs, perceptions, and learning processes during the teaching-learning process. Experience sampling methods might be appropriate (see e.g., Goetz, Frenzel, Stoeger, & Hall, 2010). Finally, cluster-randomized field trials do not allow treatment of direct or indirect effects as causal. For experimental studies to produce accurate estimates of indirect effects it is important that the intervention affect only the mediator(s) in question and that different subgroups of the experimental conditions be analyzed (Bullock, Green, & Ha, 2015). Therefore, it would seem advisable to think about experimentally manipulating mediator variables and analyzing differential effects of an intervention depending on several theoretically derived student characteristics (e.g., for learning goal orientation, see Rakoczy et al., 2013). Therefore, further research in this area should place greater emphasis on the investigation of cognitive and motivational processes explaining how formative assessment impacts learning. It should ensure the theoretical and methodological preconditions for interpreting mediating effects as causal and identify characteristics that help students detect usefulness in their teachers' feedback (moderating effects).

8.2. Final remark

At the beginning of this paper we emphasized the need to specify exactly what effective formative assessment constitutes and to investigate empirically underlying mechanisms explaining its impact on learning. We focused in our study on elicitation and feedback of diagnostic information by teachers in mathematics instruction. Accordingly, we explored students' perception of the usefulness of teachers' feedback and their self-efficacy as intervening variables for the impact of a formative assessment intervention on interest and achievement. We were able to show that both intervening variables played an important role in how formative assessment by teachers affected learning even though our results only partly met our expectations.

Appendix A. Scales

Self-efficacy

Item	Coefficient		
What do you think about the forthcoming test?	M(SD) time 1	M(SD) time 1	r_{it} time 1/ r_{it} time 2
I will probably score well on the test.	2.21 (0.61)	2.61 (0.75)	.69/.75
I believe that I can meet the requirements of this test.	2.61 (0.72)	2.96 (.70)	.69/.74
I believe that I will do well on this test.	2.36 (0.69)	2.74 (.74)	.74/.81
This test is really no problem for me.	2.32 (0.76)	2.68 (.82)	.59/.64

Interest

Item	Coefficient		
What do you think about the forthcoming test?	M(SD) time 1	M(SD) time 1	r_{it} time 1/ r_{it} time 2
The topic of the test is exciting to me.	2.31 (0.81)	2.54 (.82)	.62/.75
I like the topic of the test.	2.20 (0.78)	2.70 (.89)	.63/.75
I feel like dealing with the topic of the test.	2.44 (0.80)	2.54 (0.82)	.65/.70
The topic of the test is interesting to me.	2.32 (0.78)	2.61 (0.81)	.72/.81

[8] The effect was defined in terms of Tymms' delta measure (Tymms, 2004; Tymms, Merrell, & Henderson, 1997), which can be interpreted in line with Cohen's d, a delta of 0.5 representing a moderate effect.

Perceived usefulness

Item	Coefficient		
	M(SD) time 1	M(SD) time 2	r_{it} time 1/ r_{it} time 2
After receiving the feedback I make more effort.	3.09 (0.81)	2.84 (0.87)	.61/.56
The feedback helps me reach my learning goal.	2.93 (.67)	2.78 (0.85)	.67/.67
The feedback helps me recognize where I can improve.	3.14 (0.73)	2.88 (0.84)	.74/.73
The feedback lets me know which types of task I should practice.	3.12 (0.76)	2.89 (0.88)	.69/.68
The feedback lets me know whether I should/have to prepare myself better.	3.16 (0.78)	2.88 (0.87)	.66/.66

References

Andrade, H. L. (2010). Summing up and moving forward: Key challenges and future directions for research and development in formative assessment. In H. L. Andrade, & G. J. Cizek (Eds.). *Handbook of formative assessment* (pp. 344–351). New York, NJ: Routledge.

Asparouhov, T., & Muthén, B. O. (2010). *Multiple imputation with Mplus.* Retrieved September 29, 2010, from http://statmodel.com/download/Imputations7.pdf.

Bandura, A. (1997). *Self-efficacy: The exercise of control.* New York: Freeman.

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic and statistical considerations. *Journal of Personality and Social Psychology, 51*, 1173–1182. https://doi.org/10.1037/0022-3514.51.6.1173.

Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy & Practice, 18*(1), 5–25. https://doi.org/10.1080/0969594X.2010.513678.

Besser, M., Blum, W., & Klimczak, M. (2012). Formative assessment in every-day teaching of mathematical modelling: Implementation of written and oral feedback to competency-oriented tasks. In G. Stillman, W. Blum, J. Brown, & G. Kaiser (Eds.). *ICTMA-15 proceedings (469-478)*. New York: Springer.

Besser, M., & Leiss, D. (2014). The influence of teacher-training on in-service teachers' expertise: A teacher-training-study on formative assessment in competency-oriented mathematics. In S. Oesterle, P. Liljedahl, C. Nicol, & D. Allan (Vol. Eds.), *Proceedings of the 38th conference of the international group for the psychology of mathematics education and the 36th conference of the North American chapter of the psychology of mathematics education: Vol. 2*, (pp. 129–136). Vancouver, Canada: PME.

Black, P., & Wiliam, D. (1998a). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7–75. https://doi.org/10.1080/0969595980050102.

Black, P., & Wiliam, D. (1998b). Inside the Black Box: Raising standards through classroom assessment. *Phi Delta Kappan, 80*(2), 139–144. https://dx.doi.org10.1177/003172171009200119.

Black, P., & Wiliam, D. (1998c). *Inside the black box: Raising standards through classroom assessment.* London: Kings College, London School of Education.

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability, 21*(1), 5–31. https://doi.org/10.1007/s11092-008-9068-5.

Black, P., & Wiliam, D. (2012). Assessment for learning in the classroom. In J. Gardner (Ed.). *Assessment and learning* (pp. 11–32). (2nd ed.). London [u.a.]: SAGE.

Blum, W., & Leiss, D. (2007). Investigating quality mathematics teaching – the DISUM Project. In C. Bergsten, & B. Grevholm (Eds.). *Develobing and researching quality in mathematics teaching and learning* (pp. 3–16). Linköping: SMDF.

Briggs, D. C., Ruiz-Primo, M. A., Furtak, E., Shepard, L., & Yin, Y. (2012). Meta-analytic methodology and inferences about the efficacy of formative assessment. *Educational Measurement: Issues and Practice, 31*(4), 13–17. http://dx.doi.org/10.1111/j.1745-3992.2012.00251.

Brookhart, S. M. (1997). A theoretical framework for the role of classroom assessment in motivating student effort and achievement. *Applied Measurement in Education, 10*(2), 161–180. https://doi.org/10.1207/s15324818ame1002_4.

Bullock, J. G., Green, D. P., & Ha, S. E. (2015). Yes, but what's the mechanism? (Don't expect an easy answer). *Journal of Personality and Social Psychology, 98*, 550–558. https://doi.org/10.1037/a0018933.

Bürgermeister, A., Klieme, E., Rakoczy, K., Harks, B., & Blum, W. (2014). Formative Leistungsbeurteilung im Unterricht: Konzepte, Praxisberichte und ein neues Diagnoseinstrument für das Fach Mathematik [Formative assessment in instruction: concepts, reports, and a new diagnostic instrument for mathematics]. In M. Hasselhorn, W. Schneider, & U. Trautwein (Vol. Eds.), *Lernverlaufsdiagnostik. Tests und Trends [Formative assessment. Tests and trends]: Vol. 12*, (pp. S41–S60). Göttingen: Hogrefe.

Carver, C. S., & Scheier, M. F. (1998). *On the self-regulation of behavior.* New York, Cambridge: University Presshttp://dx.doi.org/10.1017/CBO9781139174794.

Clark, I. (2012). Formative assessment: Assessment is for self-regulated learning. *Educational Psychology Review, 24*, 205–249. http://dx.doi.org/10.1007/s10648-011-9191-6.

Cole, D. A., & Maxwell, S. E. (2003). Testing mediational models with longitudinal data: Questions and tips in the use of structural equation modeling. *Journal of Abnormal Psychology, 112*(4), 558–577. https://doi.org/10.1037/0021-843X.112.4.558.

Collins, L. M., Schafer, J. L., & Kam, C.-M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods, 6*(4), 330–351. https://doi.org/10.1037/1082-989X.6.4.330.

Dresel, M., & Ziegler, A. (2002). Failure as an element of adaptive learning. *Paper presented at the eighth biennial conference of the European association for research on adolescence*. UK: Oxford.

Dunn, K. E., & Mulvenon, S. W. (2009). A critical review of research on formative assessment: The limited scientific evidence of the impact of formative assessment in education. *Practical Assessment, Research and Evaluation, 14*(7), 1–11.

Educational Testing Service (2009). *Research rationale for the keeping learning on track program.* Princeton: Author.

Fast, L. A., Lewis, J. L., Bryant, M. J., Bocian, K. A., Cardullo, R. A., Rettig, M., et al. (2010). Does math self-efficacy mediate the effect of the perceived classroom environment in standardized math test performance? *Journal of Educational Psychology, 102*, 729–740. https://doi.org/10.1037/a0018863.

Furtak, E. M., Ruiz-Primo, M. A., Shemwell, J. T., Ayala, C. C., Brandon, P. R., Shavelson, R. J., et al. (2008). On the fidelity of implementing embedded formative assessments and its relation to student learning. *Applied Measurement in Education, 21*(4), 360–389. https://doi.org/10.1080/08957340802347852.

Gielen, S., Peeters, E., Dochy, F., Onghena, P., & Struyven, K. (2010). Improving the effectiveness of peer feedback for learning. *Learning and Instruction, 20*, 304–315. https://doi.org/10.1016/j.learninstruc.2009.08.007.

Goetz, T., Frenzel, A. C., Stoeger, H., & Hall, N. C. (2010). Antecedents of everyday positive emotions: An experience sampling analysis. *Motivation and Emotion, 34*, 49–62. https://doi.org/10.1007/s11031-009-9152-2.

Graham, J. W., Olchowski, A. E., & Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science, 8*(3), 206–213. https://doi.org/10.1007/s11121-007-0070-9.

Harks, B., Klieme, E., Hartig, J., & Leiß, D. (2014a). Separating cognitive and content domains in mathematical competence. *Educational Assessment, 19*(4), 243–266 https://doi.org/10.1080/10627197.2014.964114.

Harks, B., Rakoczy, K., Hattie, J., Besser, M., & Klieme, E. (2014b). The effects of feedback on achievement, interest and self-evaluation: The role of feedback's perceived usefulness. *Educational Psychology, 34*(4), 269–290. https://doi.org/10.1080/01443410.2013.785384.

Hattie, J. (2003). *Formative and summative interpretations of assessment information.* School of Education. The University of Auckland. Retrieved October 21, 2014, from http://web.auckland.ac.nz/uoa/fms/default/education/staff.

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

Heritage, M. (2007). Formative assessment: What do teachers need to know and do? *Phi Delta Kappan, 89*, 140–145. https://doi.org/10.1177/003172170708900210.

Hiebert, J., Gallimore, R., Garnier, H., Givvin, K. B., Hollingsworth, H., Jacobs, J., ... Stigler, J. (2003). *Teaching mathematics in seven countries: Results from the TIMSS 1999 Video Study.* NCES (2003-013), U.S. Department of EducationWashington, DC: National Center for Education Statistics.

Jiang, Y., Song, J., Lee, M., & Bong, M. (2014). Self-efficacy and achievement goals as motivational links between perceived contexts and achievement. *Educational Psychology, 34*(1), 92–117. https://doi.org/10.1080/01443410.2015.999417.

Kingston, N., & Nash, B. (2011). Formative assessment: A meta-analysis and a call for research. *Educational Measurement: Issues and Practice, 30*(4), 28–37. https://doi.org/10.1111/j.1745-3992.2011.00220.

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 Mathematikdidaktik, 31*(1), 119–141. https://doi.org/10.1007/s13138-010-0006-y.

MacKinnon, D. P., Lockwood, C. M., & Williams, J. (2004). Confidence limits for the indirect effect: Distribution of the product and resampling methods. *Multivariate Behavioral Research, 39*(1), 99–128. https://doi.org/10.1207/s15327906mbr3901_4.

McMillan, J. H., Venable, J. C., & Varier, D. (2013). Studies of the effect of formative assessment on student achievement: So much more is needed. *Practical Assessment, Research and Evaluation, 18*(2), Available online http://pareonline.net/getvn.asp?v=18&n=2.

Morin, A. J. S., Marsh, H. W., Nagengast, B., & Scalas, L. F. (2014). Doubly latent multilevel analyses of classroom climate: An illustration. *Journal of Experimental Education, 82*, 143–167. https://doi.org/10.1080/00220973.2013.769412.

Mouratidis, A., Vansteenkiste, M., & Lens, W. (2010). How you provide corrective feedback makes a difference: The motivating role of communicating in an autonomy-

supporting way. *Journal of Sport & Exercise Psychology, 32*, 619–637. https://doi.org/10.1123/jsep.32.5.619.
Muthén, L. K., & Muthén, B. O. (1998-2012). *Mplus user's guide (version 7.4)*. [Computer software]Los Angeles, CA: Muthén & Muthén.
Narciss, S. (2008). Feedback strategies for interactive learning tasks. In J. M. Spector, M. D. Merrill, J. J. G. van Merrienboer, & M. P. Driscoll (Eds.). *Handbook of research on educational communications and technology* (pp. 125–144). (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.
Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education, 31*(2), 199–218. https://doi.org/10.1080/03075070600572090.
Olkin, I., & Finn, J. D. (1995). Correlations redux. *Psychological Bulletin, 118*, 155–164. https://doi.org/10.1037/0033-2909.118.1.155.
Pajares, F. (1996). Self-efficacy beliefs in achievement settings. *Review of Educational Research, 66*, 543–578. https://doi.org/10.3102/00346543066004543.
Panadero, E., & Jonsson, A. (2013). The use of scoring rubrics for formative assessment purposes revisited: A review. *Educational Research Review, 9*, 129–144. https://doi.org/10.1016/j.edurev.2013.01.002.
Panadero, E., Jonsson, A., & Botella, J. (2017). Effects of self-assessment on self-regulated learning and self-efficacy: Four meta-analyses. *Educational Research Review, 22*, 74–98. https://doi.org/10.1016/j.edurev.2017.08.004.
Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment.* Washington, DC: National Academy Press.
Perrenoud, P. (1998). From formative evaluation to a controlled regulation of learning. Towards a wider conceptual field. *Assessment in Education: Principles, Policy & Practice, 5*(1), 85–102. https://doi.org/10.1080/0969595980050105.
Pinger, P., Rakoczy, K., Besser, M., & Klieme, E. (2016). Implementation of formative assessment –Effects of quality of program delivery on students' mathematics achievement and interest. *Assessment in Education: Principles, Policy & Practice*. Advance online publication https://doi.org/10.1080/0969594X.2016.1170665.
Preacher, K. J. (2015). Advances in mediation analysis: A survey and synthesis of new developments. *Annual Review of Psychology, 66*, 825–852. https://doi.org/10.1146/annurev-psych-010814-015258.
Preacher, K. J., & Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods, 40*, 879–891. https://doi.org/10.3758/BRM.40.3.879.
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. https://doi.org/10.1016/j.learninstruc.2013.03.002.
Rakoczy, K., Klieme, E., Leiß, D., & Blum, W. (2017). Formative assessment in mathematics in-struction: Theoretical considerations and empirical results of the Co^{2}CA project. In D. Leutner, J. Fleischer, J. Grünkorn, & E. Klieme (Eds.). *Competence assessment in education: Research, models, and instruments* (pp. 447–467). Berlin: Springer.
Rotondi, M. A. (2015). *Sample size estimation functions for cluster randomized trials (Version 1.0)*. [Software] . Retrieved December 8, 2017 from https://cran.r-project.org/web/packages/CRTSize/CRTSize.pdf.
Sadler, R. (1998). Formative assessment: Revisiting the territory. *Assessment in Education, 5*(1), 77–84. https://doi.org/10.1080/0969595980050104.
Sadler, T. D. (2011). *Socio-scientific issues in the classroom: Teaching, learning and research. Contemporary trends and issues in science education: v. 39.* Dordrecht, New York: Springer.
Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*(2), 147–177. https://doi.org/10.1037/1082-989X.7.2.147.
Schunk, D. H. (1995). Self-efficacy and education and instruction. In J. E. Maddux (Ed.). *Self-efficacy, adaptation, and adjustment: Theory, research, and applications* (pp. 281–303). New York: Plenum.
Schunk, D. H., & Pajares, F. (2009). Self-efficacy theory. In K. R. Wentzel, & A. Wigfield (Eds.). *Handbook of motivation at school* (pp. 35–53). New York: Routledge.
Schunk, D. H., & Swartz, C. W. (1993). Goals and progress feedback: Effects on self-efficacy and writing achievement. *Contemporary Educational Psychology, 18*(3), 337–354. https://doi.org/10.1006/ceps.1993.1024.
Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*(1), 153–189. https://doi.org/10.3102/0034654307313795.
Stiggins, R. (2006). Assessment for learning: A key to motivation and achievement. *Edge, 2*(2), 3–19.
Strijbos, J. W., Narciss, S., & Dünnebier, K. (2010). Peer feedback content and sender's competence level in academic writing revision tasks: Are they critical for feedback perceptions and efficiency? *Learning and Instruction, 20*, 291–303. https://doi.org/10.1016/j.learninstruc.2009.08.008.
Televantou, I., Marsh, H. W., Kyriakides, L., Nagengast, B., Fletcher, J., & Malmberg, L.-E. (2015). Phantom effects in school composition research: Consequences of failure to control biases due to measurement error in traditional multilevel models. *School Effectiveness and School Improvement, 26*, 75–101. https://doi.org/10.1080/09243453.2013.871302.
Tulis, M. (2013). Error management behavior in classrooms: Teachers' responses to students' mistakes. *Teaching and Teacher Education, 33*, 56–68. https://doi.org/10.1016/j.tate.2013.02.003.
Tulis, M., Steuer, G., & Dresel, M. (2016). Learning from errors: A model of individual processes. *Frontline Learning Research, 4*(2), 12–26. https://doi.org/10.14786/flr.v4i2.168.
Tymms, P. (2004). Effect sizes in multilevel models. In I. Schagen, & K. Elliot (Eds.). *But what does it mean? The use of effect sizes in educational research* (pp. 55–66). Slough: NFER.
Tymms, P., Merrell, C., & Henderson, B. (1997). The first year at school: A quantitative investigation of the attainment and progress of pupils. *Educational Research and Evaluation, 3*, 101–118. https://doi.org/10.1080/1380361970030201.
Van Zundert, M., Sluijsmans, D., & van Merriënboer, J. (2010). Effective peer assessment processes: Research findings and future directions. *Learning and Instruction, 20*, 270–279. http://dx.doi.org/10.1016/j.learninstruc.2009.08.004.
Wigfield, A., Eccles, J. S., & Rodriguez, D. (1998). The development of children's motivation in school contexts. *Review of research in education: Vol. 23*, (pp. 73–118). Washington DC: American Educational Research Association.
Wiliam, D., Lee, C., Harrison, C., & Black, P. (2004). Teachers developing assessment for learning: Impact on student achievement. *Assessment in Education: Principles, Policy & Practice, 11*(1), 49–65. https://doi.org/10.1080/0969594042000208994.
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.
Wilson, M. R., & Sloane, K. (2008). From principles to practice: An embedded assessment system. In W. Harlen (Ed.). *Student assessment and testing, 3* (pp. 87–112). Los Angeles, London, New Delhi, Singapore: SAGE.
Wu, M. L., Adams, R. J., & Wilson, M. R. (1998). *ConQuest: Generalised item response modelling software.* Camberwell: Australian Council for Educational Research.
Yin, Y., Shavelson, R. J., Ayala, C. C., Ruiz-Primo, M. A., Brandon, P., Furtak, E. M., et al. (2008). On the impact of formative assessment on student motivation, achievement, and conceptual change. *Applied Measurement in Education, 21*(4), 335–359. https://doi.org/10.1080/08957340802347845.