

Implementing curriculum-embedded formative assessment in primary school science classrooms

Annika Lena Hondrich, Silke Hertel, Katja Adl-Amini & Eckhard Klieme

To cite this article: Annika Lena Hondrich, Silke Hertel, Katja Adl-Amini & Eckhard Klieme (2016) Implementing curriculum-embedded formative assessment in primary school science classrooms, *Assessment in Education: Principles, Policy & Practice*, 23:3, 353-376, DOI: [10.1080/0969594X.2015.1049113](https://doi.org/10.1080/0969594X.2015.1049113)

To link to this article: <http://dx.doi.org/10.1080/0969594X.2015.1049113>



Published online: 18 Aug 2015.



Submit your article to this journal [↗](#)



Article views: 639



View related articles [↗](#)



View Crossmark data [↗](#)



Citing articles: 1 View citing articles [↗](#)

Implementing curriculum-embedded formative assessment in primary school science classrooms

Annika Lena Hondrich^{a,b*}, Silke Hertel^{b,c}, Katja Adl-Amini^{a,b} and Eckhard Klieme^{a,b}

^aGerman Institute for International Educational Research, Frankfurt, Germany; ^bCenter for Research on Individual Development and Adaptive Education of Children at Risk, Frankfurt, Germany; ^cInstitute of Educational Science, Ruprecht-Karls-University Heidelberg, Heidelberg, Germany

(Received 4 August 2014; accepted 30 April 2015)

The implementation of formative assessment strategies is challenging for teachers. We evaluated teachers' implementation fidelity of a curriculum-embedded formative assessment programme for primary school science education, investigating both material-supported, direct application and subsequent transfer. Furthermore, the relationship between implementation fidelity and teacher variables was explored. $N = 17$ German primary school teachers participated in professional development on formative assessment, $N = 11$ teachers formed a control group. Teachers' implementation fidelity was evaluated via classroom observations student ratings and an analysis of students' workbooks, focusing on the frequency and quality of intended formative assessment elements (assessments, feedback and instructional adaptations). Regarding direct application, treatment group teachers' implementation fidelity was high, with slight variations in quality. Regarding transfer, implementation fidelity was lower but teachers still implemented more formative assessment elements than the control group. Teachers' pedagogical content knowledge and their evaluation of the formative assessment intervention were associated with implementation success.

Keywords: formative assessment; fidelity of implementation; science education; teacher professional development

In recent decades, formative assessment has been widely discussed as an effective strategy for supporting students' learning (e.g. Black & Wiliam, 1998). In science education in particular, formative assessment strategies are vital to assess and react to students' preconceptions in order to help them develop a scientifically more adequate understanding (e.g. Bell & Cowie, 2001). However, research has repeatedly shown that transferring new teaching approaches into practice is not straightforward, and that it poses a major challenge for teachers to realise formative assessment strategies in instruction (e.g. Furtak et al., 2008; Smith & Gorard, 2005; Tierney, 2006). The aim of the present study is to evaluate teachers' implementation of a curriculum-embedded formative assessment programme in primary school science education. Moreover, the relationship between relevant teacher variables and implementation fidelity is explored in order to provide further insight into the processes relevant for implementing formative assessment.

*Corresponding author. Email: hondrich@dipf.de

Formative assessment

Although formative assessment has the status of a fuzzy concept (Bennett, 2011), there is agreement about its essence (e.g. Bell & Cowie, 2001; Black & Wiliam, 2009): evidence about students' understanding is elicited, interpreted and utilised in instruction with the purpose of enhancing students' learning. Following Sadler (1989), formative assessment may be conceptualised as a feedback loop: After establishing learning goals ('Where are we going?'), students' achievement is assessed in relation to these goals and their progress towards them is tracked ('Where are we now?'). To complete the formative cycle, this information is then actively used to help students proceed towards the goals ('How are we going to get there?') (Hattie & Timperley, 2007; Ruiz-Primo, Furtak, Ayala, Yin, & Shavelson, 2010, p. 139). When implemented in classrooms, the formative process can be enacted by different instructional strategies (e.g. Black & Wiliam, 2009). To use the assessment information formatively, teachers may either adapt instruction or provide feedback to students (Ruiz-Primo et al., 2010). Moreover, the degree of flexibility vs. formality may vary from gathering and using information 'on the fly', as the opportunity arises, to curriculum-embedded, preplanned assessments (Pryor & Crossouard, 2008; Shavelson et al., 2008, p. 300). These assessments are inserted at specific junctures of the curriculum, when an important learning goal should have been met. The potential of curriculum-embedded formative assessment is to provide 'thoughtful, curriculum-aligned, and valid ways of determining what students know', without 'leaving the burden of planning and assessing on the teacher alone' (Shavelson et al., 2008, p. 3).

Formative assessment strategies have been applied in many contexts, including science education. Research has shown that primary school students enter the classroom with their own explanations for scientific phenomena, which are often functional in everyday life but more or less incompatible with scientific conceptions (Hardy, Jonen, Möller, & Stern, 2006; Morrison & Lederman, 2003). These preconceptions or misconceptions are usually deep-rooted and need to be specifically addressed in instruction to be modified (Hardy et al., 2006; Vosniadou, 2008). Therefore, it seems particularly promising to use formative strategies in early science classrooms to explore students' scientific reasoning, evaluate misconceptions and respond to them in order to enable conceptual development (Bell & Cowie, 2001). This view is supported by a number of empirical studies providing evidence that formative assessment can indeed be effective in promoting students' learning and conceptual change (Kingston & Nash, 2011; Tomita, 2009).

However, successful implementation of formative assessment requires not only that the steps of the formative cycle are enacted, but also that quality features are met (Furtak et al., 2008). Assessments need to address students' underlying reasoning and tap on relevant conceptions and misconceptions in order to provide valid information on students' learning. Moreover, it is essential that formative feedback focuses on the levels of task solution and learning processes, providing students with specific information on their current understanding as well as with strategies about how to take the next steps in learning (Hattie & Timperley, 2007; Ruiz-Primo et al., 2010). Finally, one important goal of formative assessment is to encourage students to take responsibility for their own learning – by making the formative process transparent to students and creating a constructive learning atmosphere in which mistakes are regarded as valuable information (Black & Wiliam, 2009). For example, feedback

has a positive impact on learning only if students understand, accept and use the information to increase effort or adjust their learning strategies (Hattie & Timperley, 2007). High-quality formative assessment thus shares some basic ideas with concepts of general teaching quality – for example, constructive formative feedback is also an aspect of ‘constructive climate’ which is often considered as a dimension of teaching quality (Klieme, Pauli, & Reusser, 2009; for a detailed discussion, see Decristan et al., *in press*).

Implementing formative assessment in classroom instruction

When considering the requirements for high-quality formative assessment, it is not surprising that its realisation poses considerable challenges for teachers (e.g. Yin et al., 2008). Indeed, research indicates that formative assessment strategies are still not being used systematically in classroom instruction and even if, essential quality features are often not regarded (e.g. Morrison & Lederman, 2003; Noyce, 2011). Teacher professional development programmes are a potentially effective means of fostering teachers’ use of formative assessment strategies (e.g. Torrance & Pryor, 2001; Wiliam, Lee, Harrison, & Black, 2004), but still, teachers have repeatedly experienced problems in implementing formative assessment strategies in intervention studies (e.g. Furtak et al., 2008). Thus, it is crucial to evaluate whether and to what extent teachers realise formative assessment programmes in their classrooms, and which factors may foster or hinder successful implementation.

Implementation fidelity

Implementation fidelity refers to the extent to which teachers’ *enacted* classroom practice reflects an *intended treatment*, for example a programme conveyed in a professional development workshop (Dusenbury, Brannigan, Falco, & Hansen, 2003; O’Donnell, 2008). There are different approaches to conceptualising and assessing implementation fidelity (Mowbray, Holter, Teague, & Bybee, 2003). Crucially, the critical components of the treatment that are expected to convey its positive effects must be evaluated, considering their quantity as well as quality (Gresham, 2009; Ruiz-Primo, 2006). Regarding curriculum-embedded formative assessment, implementation fidelity therefore includes the usage of embedded formative assessment elements (e.g. provision of feedback) as well as the quality of these elements and their enactment (e.g. feedback including strategies for improvement). Implementation fidelity may be evaluated from the observers’, the teachers’ and the students’ perspective. The observers’ perspective is generally regarded as an objective and valid source of information (e.g. Dusenbury et al., 2003). However, the impact of formative assessment strategies also depends on the students’ perception of feedback and learning climate as constructive. Hence, the students’ perception of these processes – also called their ‘responsiveness’ to the intervention (Dusenbury et al., 2003) – is an important quality component of formative assessment implementation.

Factors influencing the implementation of formative assessment

Research has shown that in general, several factors influence teachers’ implementation fidelity of new approaches (Desimone, 2009; Penuel, Fishman, Yamaguchi, & Gallagher, 2007). Among them are teacher variables (e.g. motivation, knowledge

base, and prior beliefs), aspects of the professional development intervention (duration, characteristics of the programme) as well as context variables (class characteristics, support from the environment). For example, research shows that professional development is most successful when it starts out at teachers' interests and needs, leaving room for their active, practice-oriented engagement (Desimone, 2009; Postholm, 2012). Moreover, implementation fidelity is positively influenced if support is provided through adequate training and manuals; it is likely to be lower if the programme is complex, time-consuming and requires additional materials (Gresham, 1989).

Although systematic research investigating these factors in the implementation of formative assessment is scarce, the existing evidence, mainly from qualitative studies, suggests that they are also relevant for the implementation of formative assessment (e.g. Dixon, Hawe, & Parr, 2011; Tierney, 2006). Analysing the requirements for implementing high-quality formative assessment, two major issues can be identified. First, formative assessment makes high demands on teachers' pedagogical content knowledge (Ruiz-Primo et al., 2010; Shulman, 1986). This includes a clear conception of learning goals and the awareness of typical misconceptions or learning problems; the ability to adequately assess them and finally, a set of suitable strategies for moving students forward in their learning. Thus, teachers' pedagogical content knowledge is central for their ability to realise high-quality formative assessment (see also Falk, 2011). Moreover, a profound knowledge about students' potential misconceptions should affect teachers' motivation to use formative assessment, as it makes evident the importance to address students' misconceptions within the learning process. The fact that the realisation of formative assessment is always content-specific also indicates particular challenges regarding the transfer of formative assessment strategies: for each curricular content, specific knowledge and materials are required (Bennett, 2011; Coffey, Hammer, Levin, & Grant, 2011).

The second major barrier for realising formative assessment strategies is the time and effort needed for preparation and application, especially when the method is newly implemented (e.g. Lee, Feldman, & Beatty, 2011; Torrance & Pryor, 2001) – for example, delivering formative feedback to each student is more time-consuming than grading. Teachers therefore need effective strategies for realising formative assessment within the given time constraints. Furthermore, teachers' motivation to use formative assessment is vital in this context. Their beliefs concerning the relevance and effectiveness of formative assessment are therefore considered central for the realisation of formative assessment strategies (e.g. Ruiz-Primo et al., 2010). Indeed, in a review by Tierney (2006), teachers' educational beliefs as well as time concerns were identified as influencing factors for the implementation of formative assessment.

It is important to address these challenges within professional development when supporting teachers to realise formative assessment in their classrooms (Wiliam et al., 2004), as well as to investigate the actual impact of these factors on teachers' implementation of formative assessment strategies. This applies particularly to primary school science class. Research shows that primary school teachers often lack pedagogical content knowledge of scientific topics (Appleton, 2007) and display low self-efficacy in teaching science (Watters & Ginns, 1997). Thus, we can expect many teachers to face particular difficulties in realising formative assessment in this context. Therefore, more research is needed on the implementation of formative assessment in primary school science education.

Research aims

Research aim 1: teachers' implementation fidelity

In the present study, our central research aim is to evaluate the implementation success of a curriculum-embedded formative assessment intervention in primary school science classrooms. We evaluate teachers' implementation fidelity in two implementation conditions, i.e. a direct application condition in which teachers' implementation is supported by pre-designed materials, and a subsequent transfer condition in which the teachers are encouraged to design the formative assessment materials on their own. Teachers' implementation fidelity is evaluated assessing the frequency and quality of formative assessment elements and their enactment, basing on external observations and students' ratings. The following hypotheses were specified:

- (1) In the direct application and transfer condition, teachers participating in the formative assessment intervention will show more curriculum-embedded formative assessment strategies than an untreated control group.
- (2) When comparing direct application against transfer, treatment group teachers' implementation fidelity in the transfer condition will be lower and show more inter-individual variation.

Research aim 2: variables connected to teachers' implementation fidelity

Our second research aim is to evaluate the relationship between teachers' implementation fidelity and specific variables on teacher level. As teacher correlates, we investigated teachers' knowledge of students' misconceptions – a central aspect of pedagogical content knowledge – and teachers' evaluation of the formative assessment programme. The following hypotheses were specified:

- (1) The more teachers know about students' misconceptions, the more often will they implement formative assessment elements and the higher will the quality of implementation be, resulting in a higher implementation fidelity in both direct application and transfer condition.
- (2) The more positively teachers evaluate the formative assessment programme after the direct application condition, the more successful they will be in implementing the programme in the transfer condition.

Methods

Design

The present study was part of a research initiative which set out to evaluate different teaching strategies for inquiry-based science education (Decristan et al., 2015; Hardy et al., 2011). Teachers in the study were randomly assigned on school level to two conditions, *formative assessment (FA)* and *control (CG)*. Teachers of both groups took part in professional development workshops on the topic of floating and sinking, including pedagogical content knowledge. The FA group additionally received training in formative assessment. All teachers then taught two curriculum units on floating and sinking in their classrooms, FA group teachers were asked to additionally use formative assessment strategies. In the first unit, they were supported by formative assessment materials and a manual (direct application condition). In the

second unit, they were encouraged to design the formative assessment materials on their own based on the principles presented in the professional development workshops and following the examples of the first unit (transfer condition). Each unit consisted of 9 lessons lasting 45 min each, combinable as double lessons, and was expected to span slightly more than two weeks.

Sample

The sample underlying the present study consists of $N = 28$ German primary school teachers from 18 schools. Each teacher participated with his or her third grade science class (in all, $N = 519$ students). 17 teachers participated in the formative assessment condition ($n = 319$ students) and 11 teachers in the control condition ($n = 200$ students). Class size varied between 10 and 26 students, with a mean of 19.7 students. At the beginning of the study, the participating students were on average 8.8 years old ($SD = .5$) with a proportion of 52.1% males.

The participating teachers were mostly female (85.7%), had a mean age of 43.4 years ($SD = 9.8$) and an average teaching experience of 15.8 years ($SD = 9.8$). While all teachers had taught science within the past five years, only three had received professional training, one of whom in the FA group. There were no significant differences between the two groups regarding age, gender, teachers' experience and their professional training ($p \geq .54$, using the t -test and the chi-square test). However, FA classes tended to be slightly smaller than control classes ($t = 1.819$, $p = .08$).

Treatment

The curriculum

Our study was based on two third grade units on floating and sinking adapted from Jonen and Möller (2005). The overarching learning goal of the first unit was to understand and apply the concept of relative density, which was subdivided into four learning steps: (1) disproving common misconceptions on floating and sinking; (2) floating and sinking as a property of material; (3) simple density of materials; and finally, (4) using relative density to predict the floating or sinking of objects. The second unit focused on the concepts of buoyancy force and displacement in order to build an integrated conception of floating and sinking, and had a comparable structure with four learning steps. Within both units, each step was implemented using an inquiry-based approach (e.g. Anderson, 2002), allowing students to construct knowledge through a process of active scientific investigation and evaluation of empirical evidence. Starting with a research question, students' hypotheses were collected and experiments were planned, conducted and discussed. Finally, the findings were applied using differentiated worksheet tasks. Following theories of 'intelligent training' (Helmke, 2006), teachers could choose tasks from three levels: complex transfer tasks, consolidation tasks and basic repetition tasks (often involving additional student experiments challenging specific misconceptions).


Curriculum-embedded formative assessment

Our programme of curriculum-embedded formative assessment included three main elements: (a) short written tasks to assess students' current conceptual

understanding, (b) individual, written, semi-standardised feedback and (c) the adaptation of instruction by assigning differentiated worksheet tasks based on assessment information. When implementing the formative assessment elements in their classrooms, teachers were asked to emphasise the formative purpose of the assessments and feedback and frame them with the students' activity as 'researchers' who constantly probe and revise their ideas to improve their understanding.

Diagnostic assessments. The assessments were developed for the study, partly adapted from Jonen and Möller (2005). In the sense of a two-tier formative assessment probe (Keeley, 2005), we used a combination of open as well as multiple-choice answer formats to assess students' conceptions on floating and sinking. Figure 1 gives an example of the assessments used in the study. The assessments (four in all) were embedded after each learning step and at the end of a lesson, so that teachers could evaluate students' answers after school and use them for preparing the next lesson. All assessments addressed the conceptions on floating and sinking students argued with. Additionally, assessments 2, 3, and 4 assessed how well students applied the conception introduced in the respective lesson. We provided teachers with a guideline on how to interpret students' results (see Appendix 1) as well as with a table for documenting students' conceptions and levels of understanding. Students were classified according to three levels of understanding: those students who could apply the new concept very well, mastering transfer tasks and reliably rejecting misconceptions (level 3); those who showed some understanding but still made mistakes in complex tasks triggering misconceptions (level 2); and those who

Brain teaser No. 4



1. Which cubes float?

a) Here are some more cubes. They are just as big as our "water cube". Check the correct box: do the cubes float or sink in water?





<p>water cube</p>  <p>61g</p>		
<p>brown plastic cube</p>  <p>49g</p>	<p>red plastic cube</p>  <p>62g</p>	<p>clay cube</p>  <p>93g</p>
<input type="checkbox"/> floats <input type="checkbox"/> sinks	<input type="checkbox"/> floats <input type="checkbox"/> sinks	<input type="checkbox"/> floats <input type="checkbox"/> sinks

Figure 1. Excerpt from assessment No. 4, evaluating students' understanding of relative density (translated from German).

were still unable to apply the new concept even in simple tasks (level 1). These levels constituted the basis for the further formative usage of the information that had been gathered.

Formative feedback. The assessment information was used to provide written feedback to students. Drawing on research on effective formative feedback (Hattie & Timperley, 2007), the feedback should (a) inform students about how well they had understood the targeted concept, including knowledge of response and feedback on specific problems or misconceptions, if present; and (b) provide students with a strategy for improvement when working on the next tasks. Teachers were instructed to provide complete formative feedback twice (after assessments 2 and 4) and were asked to indicate knowledge of response after assessment 3. In order to help teachers realise formative feedback as intended, we provided teachers with feedback templates for the three levels of students' understanding. Teachers needed to fill in specific problems the students had faced and were encouraged to add additional, differentiating information whenever necessary.

Adaptive instruction. Moreover, the assessment information was used to adapt instruction to students' level of understanding. The first assessment, focusing on students' preconceptions, served teachers to prepare classroom discussions and experimental tasks. After assessments 2, 3 and 4, teachers were instructed to assign the available differentiated tasks according to the three levels of students' understanding: complex transfer tasks were given to students who had reached level 3; consolidation tasks were assigned to level 2 students, and basic repetition tasks were given to level 1 students.

Professional development workshops

Five professional development workshops (each taking 4.5 h) were held for both the treatment and control group. The workshops were designed and given by staff members, with valuable input from practitioners in all phases of development. In all workshops, room was provided for hands-on activities and discussion among the participants as principles of successful professional development (Desimone, 2009). Two workshops - held by the same training team for both groups - addressed the curriculum on floating and sinking, including pedagogical content knowledge on density (workshop 1), as well as on buoyancy force and displacement (workshop 5). In between, teachers in the formative assessment group attended three workshops on formative assessment, while control group teachers took part in three workshops on parental counselling instead. In the treatment group, workshop 2 focused on the concept of formative assessment and its impact on students' learning and motivation. Workshop 3 dealt with implementing formative assessment within the first curriculum unit, familiarising teachers with the operationalization of formative assessment as embedded in the curriculum. Finally, workshop 4 focused on realising formative assessment in teaching practice and transfer to other topics.

All teachers received standardised materials (ranging from worksheets to materials for experiments) and a detailed manual for teaching the curriculum. For the first unit, FA teachers received a formative assessment version of the manual which additionally included the formative assessment materials described above. The first unit was taught after workshop 4, the second unit after workshop 5.

Implementation fidelity of formative assessment

Implementation fidelity was primarily evaluated by external observations, complemented by students' ratings. First, we conducted classroom observations of a double lesson (90 min) for all teachers in the sample – either video-based ($n = 20$) or live for those teachers and classes who did not agree to be filmed ($n = 8$). Second, within the treatment group only, we analysed students' workbooks with their collection of all didactical materials used throughout the units. As third source of information, we used students' ratings of perceived formative assessment strategies in both groups (see below). The observation scores were based on dichotomous items checking the occurrence of a certain element or quality aspect. All observation items had good inter-rater agreement ($\geq 85\%$); we could not compute Cohen's κ due to a lack of variance in several items (element realised by all teachers, as observed by all raters).

Frequency of curriculum-embedded formative assessment elements

Implementation frequency was evaluated by external observations. Basing on students' workbooks, we evaluated the occurrence of written assessment, feedback and adaptive instruction throughout each unit, using 12 dichotomous items (see Appendix, Table A1). Four items referred to the occurrence of embedded assessment on the contents of lesson 1–4, four items referred to the provision of individual, written feedback, and four items referred to adaptive instruction assessing the assignment of differentiated tasks. The implementation score for each component was computed as the percentage of enacted formative elements relative to the maximum number of elements as intended by the manual. The mean percentage of the three components formed the overall frequency score. In the control group, we did not have access to students' materials. Therefore, we used the classroom observations of one double lesson to compare the frequency of curriculum-embedded formative assessment elements in the treatment and the control group.¹ For the observed lesson, we formed a score parallel to the frequency score, evaluating the implementation of assessment feedback and adaptation of instruction (four items in all). Percentage scores were computed according to the intended number of elements given in the FA treatment manual.

Quality of formative assessment implementation

To evaluate the quality of implementation in the treatment group, we assessed the occurrence or non-occurrence of basic quality aspects in assessments, feedback and their enactment in the classroom. We did not assess the quality of adaptive instruction, as teachers all used the same differentiated tasks in both units with no variations in quality. Moreover, we used students' ratings of teachers' formative assessment practices as quality indicator in both groups.

Quality of assessments. While teachers could use pre-designed assessments in the direct application condition, they had to design their own assessment tasks in transfer – with potentially varying quality. As quality indicators, we evaluated if the assessments included tasks assessing students' knowledge of the target concept(s) (as taught in the preceding lesson) and included tasks suitable for assessing relevant misconceptions (open questions requiring explanations or tasks directly triggering

misconceptions). Thus, the maximum score for each assessment was 2. A mean assessment quality score was computed across up to four assessments (as planned in the manual).

Quality of feedback. When providing feedback, teachers were asked to verify students' responses, inform about their understanding relative to the learning goal and indicate a strategy for improvement. We used three dichotomous items to rate if these components were present, so that the maximum score was 3. A mean feedback quality score was computed across up to two 'complete' feedbacks as planned in the manual.

Quality of enactment – transparency. We evaluated the quality of enactment of formative assessment elements, focusing on transparency: did teachers make the formative purpose explicit to students? Based on classroom observations, we included two dichotomous items: 'teachers discussing assessments and/or feedback with the class' and 'teachers explicitly stating the formative purpose of the assessments'. Teachers' maximum score was thus 2.

Students' perception of formative assessment strategies in treatment and control group

As additional quality component of implementation fidelity, we assessed students' perception of teachers' formative assessment strategies in both the treatment and control group, before the intervention and after each unit. We used a rating scale with eight items which was adapted from Fauth, Decristan, Rieser, Klieme, and Büttner (2014) and covers aspects of teaching quality that are relevant to the quality of formative assessment implementation. Items referred to strategies like informative, constructive feedback and ongoing assessment of students' understanding (e.g. 'My science teacher keeps checking what I already know and what I still need to learn'). Items were rated on a four point-Likert scale (1 'strongly disagree' – 4 'strongly agree'). As we were interested in students' shared perception of teacher behaviour, their mean scores were aggregated on class level. Reliability was good ($ICC2 \geq .76$). Prior to the intervention, no difference in students ratings was found between the FA and control group ($t = -.085$; $p = .93$).

Teacher measures

After workshop 4, immediately before formative assessment was enacted for the first time, FA group teachers' knowledge of students' misconceptions concerning floating and sinking was assessed as an aspect of their pedagogical content knowledge. In an open answer format, teachers were asked to name all misconceptions of students which might influence their learning process. Answers were coded by trained staff to identify the sum of relevant misconceptions provided; inter-rater agreement was high ($ICC = .95$). Due to missing values, data was available for $n = 14$ teachers.

FA group teachers' evaluation of the formative assessment programme was assessed after the first unit using a 20 item scale (Cronbach's $\alpha = .81$). Items referred to the quality of provided materials and the perceived impact on students' learning and motivation (e.g. 'the feedback fostered students' learning'). Items were rated on a four point-Likert scale (1 'strongly disagree' – 4 'strongly agree').

Results

Research aim 1: teachers' implementation fidelity

Our first research aim was to describe and evaluate teachers' implementation fidelity in a direct application and a transfer condition.

Comparison between the treatment group and control group

We expected teachers in the treatment group to implement more curriculum-embedded formative assessment elements than an untreated control group in both direct application and transfer condition. Indeed, as rated by observers, in the observed lessons, control group teachers used none of the curriculum-embedded formative assessment strategies planned in the intervention – written diagnostic tasks, written individual feedback, assigning specific tasks to students – of their own accord (see Table 1). Treatment group teachers, by contrast, showed a high frequency of curriculum-embedded formative assessment in the direct application condition. In the transfer condition, implementation was scarcer, but still about a quarter of the planned elements was implemented. The differences between treatment group and control group were significant in both conditions, tested with the Mann–Whitney *U*-test due to non-normal distribution of variables ($z \leq 3.13$; $p < .01$; $d \geq 1.31$).

Moreover, we expected teachers' implementation of formative assessment strategies to be perceived by students. Despite rather high reported levels of perceived formative assessment strategies in all classes (see Table 1), scores in the FA group were significantly higher than in the control group in both conditions (tested with the *t*-test; direct application: $t \leq 1.72$; $p < .05$; $d = .66$; transfer: $t \leq 2.04$; $p = .03$; $d = .83$).

Implementation fidelity in the treatment group: comparison of direct application and transfer

For comparing the two conditions, we focused on the treatment group, analysing the implementation frequency and quality in more detail.

Implementation frequency. Teachers' implementation fidelity of the formative assessment intervention in the direct application condition was high throughout the unit (see Table 2 for descriptive measures). As the analysis of students' materials

Table 1. Comparison of implementation fidelity in the FA group vs. control group (observed implementation frequency and students' perception of formative assessment (FA) strategies).

	Direct application		Transfer	
	CG	FA	CG	FA
Frequency of intended FA elements (in percent) – classroom observations <i>M</i> (SD)	.00 (.00)	95.59 (13.21)	.00 (.00)	27.94 (27.79)
Students' perception of FA strategies (class means) <i>M</i> (SD)	3.15 (.34)	3.34 (.27)	3.07 (.39)	3.34 (.30)

Note: Frequency score based on classroom observations (one lesson).

showed, all teachers implemented the assessments as intended. The (rather small) differences among teachers originate in their different implementation of feedback and adaptive instruction. Only one teacher (teacher 3) showed a lower implementation frequency of 55.56%. This teacher had been unable to take part in the professional development workshops 1–4 but had received all supportive materials.

Regarding the transfer condition, FA group teachers on average implemented about a third of the intended formative assessment elements throughout the unit, but showed considerable inter-individual variation. Five of the 17 teachers showed none of the intended elements, while the other twelve used varying amounts of formative assessment (see Table 2).

Comparing implementation frequency between both conditions, the Wilcoxon-test (used due to non-normality of variables) showed that teachers implemented significantly fewer formative assessment elements in the transfer condition than in the direct application condition ($z = 3.58$; $p < .001$; $d = 2.87$). Comparing the variance of implementation frequency with the Pitman's test (modified by Grambsch, 1994), we found a significantly higher variability in the transfer condition ($z = 2.35$; $p = .02$).

Implementation quality. In the direct application condition, all teachers used the pre-designed assessments, which fulfilled the quality requirements (see Table 2). Feedback quality was also high. However, when enacting feedback and assessments in the observed lesson, teachers showed less transparency than intended. In the transfer condition, feedback and assessment quality could only be evaluated for those teachers who had actually used the respective elements, and the transparency score only if they had also shown them in the observed lesson. Assessment quality in transfer was almost as high as in direct application and showed low inter-individual variance, indicating that most teachers addressed the relevant conceptions and misconceptions in their self-designed assessments. However, the feedback provided showed considerably lower quality, and teachers' mean transparency score dropped even further compared to the direct application condition. The differences in feedback quality and transparency were significant between conditions ($z \geq 2.45$; $p \leq .02$; $d \geq 1.19$), showing that teachers' feedback and enactment of formative assessment quality were indeed lower in the transfer condition. Regarding assessment quality, the difference did not reach statistical significance ($z = 1.63$; $p = .05$).

Students' perception of formative assessment strategies. Students rated their teachers' use of formative assessment strategies exactly the same in both direct application and transfer (see Tables 1 and 2). Thus, the expected difference in implementation fidelity between the two conditions was not present in students' judgments. Although the variance between classes was slightly higher in transfer, the difference was not significant (F -test; $F = 1.27$, $p = .32$).

Research aim 2: relationship between teacher variables and implementation fidelity

As our second research aim, we investigated correlations of FA group teachers' implementation fidelity (as assessed by the material-based frequency score and the quality scores) with their knowledge of student misconceptions as well as with their

Table 2. Implementation fidelity: treatment group teachers' implementation frequency (percentage scores) and implementation quality, rated by observers and students.

	Direct application			Transfer		
	<i>M</i>	SD	Range	<i>M</i>	SD	Range
Frequency – overall	95.43	11.15	55.56–100.00	33.50	31.53	.00–111.11 ^b
Frequency – assessment	100.00	.00	100.00–100.00	33.82	29.24	.00–100.00
Frequency – feedback	92.16	14.58	66.67–100.00	41.18	38.24	.00–133.33 ^b
Frequency – adaptive instruction	94.12	24.25	.00–100.00	25.49	32.34	.00–100.00
Quality of assessment	2.00	.00	2.00–2.00	1.89 ^a (<i>n</i> = 12)	.21	1.50–2.00
Quality of feedback	2.56	.64	1.00–3.00	1.50 ^a (<i>n</i> = 11)	.77	.50–3.00
Quality of enactment – transparency	1.24	.56	.00–2.00	.64 ^a (<i>n</i> = 8)	.52	.00–1.00
Students' perception of FA strategies (class means)	3.34	.27	2.85–3.75	3.34	.30	2.68–3.79

Note: Frequency scores based on documentary analysis of students' materials.

^aQuality scores were computed only for teachers who implemented the respective elements at least once (and, for transparency, in the observed lesson).

^bScores higher than 100% indicate that teachers performed more elements than proposed in the manual.

evaluation of the formative assessment programme. As teacher 3 could not participate in four of the five professional development sessions, this teacher was omitted from the following analyses. Moreover, the quality of assessment scale showed reduced variance between teachers, so that correlations could either not be computed (direct application condition) or should be interpreted with care (transfer condition).

Table 3 shows correlations of teacher variables with teachers' implementation fidelity. Teachers on average displayed knowledge of *M* = 4 potential misconceptions of students in explaining floating and sinking (*SD* = 1.33). Knowledge of students' misconceptions (MIS) was positively correlated with all implementation fidelity scores in direct application. In transfer, we only found a high, significant correlation with teachers' transparency scores (notably, given the low number of cases, this result should be interpreted with caution).

Teachers' evaluation of the formative assessment programme (EVA) after the direct application condition was predominantly positive: on average, they evaluated the formative assessment elements and materials as 'rather apt' for fostering students' learning and motivation (*M* = 2.99, *SD* = .34). As expected, EVA was positively correlated with teachers' implementation frequency in both direct application and transfer condition. Moreover, we found significant positive correlations with the quality indicators transparency and students' perception of formative assessment strategies in direct application, and with feedback quality in the transfer condition.

Discussion

In our study, we set out to implement curriculum-embedded formative assessment in primary school science classrooms. Our central research aim was to evaluate

Table 3. Correlations (Spearman's ρ) of implementation fidelity scores with teachers' knowledge of students' misconceptions (MIS) and their evaluation of the formative assessment concept (EVA).

	MIS		EVA ^a	
	Teachers' knowledge of misconceptions		Evaluation of the formative assessment concept	
	Direct application	Transfer	Direct application	Transfer
Frequency – overall ^b	.58* (<i>n</i> = 14)	.28 (<i>n</i> = 14)	.44* (<i>n</i> = 16)	.78** (<i>n</i> = 16)
Quality of assessment	– ^c	.38 (<i>n</i> = 10)	– ^c	–.44 (<i>n</i> = 12)
Quality of feedback	.52* (<i>n</i> = 14)	.27 (<i>n</i> = 9)	.35 (<i>n</i> = 16)	.56* (<i>n</i> = 11)
Quality of enactment – transparency	.55* (<i>n</i> = 14)	.88** (<i>n</i> = 7)	.48* (<i>n</i> = 16)	.22 (<i>n</i> = 8)
Students' perception of FA strategies (class means)	.50* (<i>n</i> = 14)	.43 ⁺ (<i>n</i> = 14)	.56* (<i>n</i> = 16)	.18 (<i>n</i> = 16)

^aEVA was assessed after the direct application condition.

^bCorrelations with frequency subscores are not reported, as they are highly correlated with the over-all score and lack variance in the direct application condition.

^cNo variance in Quality of assessment.

⁺ $p < .10$; * $p < .05$; ** $p < .01$ (one-tailed).

teachers' implementation fidelity – both in a material-supported, direct application condition and in a subsequent transfer condition. Moreover, we examined the relationship of teachers' implementation fidelity with their knowledge of student misconceptions and their evaluation of the formative assessment programme.

Research aim 1: teachers' implementation fidelity

Comparing treatment group teachers' implementation of formative assessment strategies with the control group via observers' ratings showed that, as expected, the treatment enabled teachers to use significantly more formative assessment elements than an untreated control group in both direct application and transfer. Indeed, no teacher from the control group used anything similar to the proposed curriculum-embedded formative assessment strategies (written diagnostic task, written individual feedback, individual assignment of tasks) in the observed lesson. This supports findings that particularly written and pre-planned formative assessment strategies are used rarely and unsystematically in regular classroom instruction (Morrison & Lederman, 2003) and that professional development can be a successful means of fostering teachers' use of formative assessment strategies (e.g. Wiliam et al., 2004). Moreover, the difference between treatment and control group was also evident in students' ratings, as FA students reported higher levels of formative assessment strategies in direct application and transfer. This shows that students perceived relevant changes in teachers' behaviour following the formative assessment intervention. However, one has to note that students' ratings were rather high in both groups, which may be due to the fact that the included strategies are less specific than the observed elements (for example, 'repeatedly checking' students' understanding can not only be realised by written assessments, but also through oral questioning).

In the following step, we analysed implementation fidelity in the treatment group in more detail, contrasting direct application with transfer. In the direct application condition, implementation fidelity was high. Analysis of students' workbooks showed that almost all teachers showed near perfect implementation frequency. Moreover, teachers generally provided high quality feedback, informing students about their current understanding and strategies for improvement. Only the observed transparency of the formative process was less well enacted, meaning that some teachers did not explicate the formative aim of assessments and feedback to their students. These results indicate that primary school teachers are able to implement most aspects of a curriculum-embedded formative assessment intervention when it is combined with supportive materials and professional development workshops – even given a challenging subject like science. This is not self-evident as although teachers were provided with all the necessary tools, it still takes time to evaluate students' assessments, document their progress and apply the provided feedback templates to each individual student. In contrast to their implementation success in direct application, most teachers had substantial problems in transferring the formative assessment programme to another topic – meaning that they had to design and evaluate their own assessments, provide elaborate feedback and assign appropriate tasks. Presumably, devising the necessary materials proved too difficult or time-consuming for teachers to keep up with the intended rate of four assessments within two weeks. Moreover, feedback quality and the transparency of enactment were significantly lower than in the direct application condition. Only in assessment quality and students' ratings, no difference was found between the conditions.

These results underline the importance of supportive materials and explicit training in professional development workshops for teachers' implementation of formative assessment, confirming previous findings (e.g. Desimone, 2009; Gresham, 1989). As a consequence, teachers either need a large database of suitable formative assessment materials or should be offered more explicit support in developing transferable strategies. Research suggests that a promising approach is to leave room for teachers to develop and explore their own formative assessment tools, accompanied by peer as well as trainer support. This takes more time but may facilitate a flexible use of formative assessment across curricula and classes (Postholm, 2012; Wiliam et al., 2004). Moreover, the aspect of transparency of the formative process obviously needs further support. Implementing transparency is a more fundamental process than using predesigned formative assessment elements because it is associated with a shift from a teacher-centred assessment culture to the inclusion of students in this process, with shared responsibility for learning success (Black & Wiliam, 2009). Professional development should focus more strongly on this issue and point to strategies like the sharing of evaluation criteria.

It is remarkable that students' ratings of formative assessment strategies were equally high in transfer despite lower levels of implementation frequency and quality of formative assessment elements. This gives first hints that the intervention may have caused changes in teachers' behaviour even for those teachers who reduced or omitted the intended elements, for example by placing more emphasis on oral, on-the-fly formative assessment strategies. Still, the conclusion that teachers transferred the principles of formative assessment well and simply changed their tools is premature, as then we would also expect higher levels of implementation quality, particularly transparency. Another explanation is that even a rare use of formative

assessment strategies was salient to students as compared to their other teaching experiences, leading to high ratings.

Finally, it is important to note the substantial inter-individual variation of implementation fidelity in the transfer condition. This indicates that – as we investigated in our second research aim – differing preconditions, for example on teacher level, influence implementation fidelity, especially when little support is provided.

Research aim 2: variables connected to teachers' implementation fidelity

The results regarding our second research aim show that both teachers' evaluation of the formative assessment programme after the first unit (EVA) and their knowledge of relevant misconceptions (MIS) were correlated with measures of implementation fidelity.

EVA was strongly and significantly correlated with teachers' implementation frequency; correlations with quality scores were less consistent but present, for example for feedback quality in the transfer condition. These results indicate that the more teachers were satisfied with the programme and the exemplary materials after using them in classroom instruction, the more they invested in the transfer of the programme to another topic. This is in line with earlier findings that the perceived effectiveness of a programme is central for its successful implementation (e.g. Desimone, 2009; Gresham, 1989). Our results also suggest that the same intervention programme may be well suited for some teachers, while it seems to be less effective for others. As practical implication, formative assessment programmes should be carefully evaluated, and adapted if necessary to gain teachers' acceptance. Moreover, teacher professional development should offer possibilities for teachers to flexibly integrate existing formative assessment programmes into their teaching routines (Wiliam et al., 2004).

Another important variable that may influence teachers' implementation of formative assessment is teachers' knowledge of students' misconceptions (MIS). As expected, MIS was significantly correlated with teachers' implementation fidelity scores in the direct application condition. In the transfer condition, MIS was highly and significantly correlated with the quality of enactment (transparency). These results confirm the importance of pedagogical content knowledge for the implementation of formative assessment (e.g. Ruiz-Primo et al., 2010), especially for a transparent enactment of formative assessment. This indicates that addressing pedagogical content knowledge within professional development on formative assessment is indeed a useful strategy to support implementation fidelity. However, contrary to expectations and to previous research (Falk, 2011; Ruiz-Primo et al., 2010), teachers' knowledge of misconceptions was uncorrelated with implementation frequency and the remaining quality scores in the transfer condition. One possible explanation lies within the design of the study: Teachers' knowledge of students' misconceptions was assessed immediately before the direct application condition, but still a few weeks ahead of the transfer unit. Teachers' knowledge of student misconceptions may have changed owing to workshop session 5 and their experiences with the first unit.

Limitations and implications for further research

The present study has several limitations which need to be addressed. Among them is the small sample size of $n = 11$ teachers in the control group and $n = 17$ teachers

in the treatment group, which was further reduced in the analysis of implementation quality. Therefore, although teachers were randomly assigned to the formative assessment intervention, results should only be generalised with caution. Although small sample sizes are common in formative assessment intervention studies (e.g. Kingston & Nash, 2011), in the future, researchers should make efforts to scale up interventions.

As we aimed at implementing an existing programme, we had a clear vision of what teachers were expected to implement as formative assessment (*intended treatment*). Thus, the implementation fidelity scores focus primarily on curriculum-embedded formative assessment rather than assessing if and to what extent teachers used other forms, like on-the-fly formative assessment (Shavelson et al., 2008). Only the students' ratings refer to general strategies not limited to embedded, written forms of formative assessment. Therefore, it would be interesting to investigate in more detail and from an observers' perspective whether other forms of formative assessment were also increasingly used by teachers from the treatment group compared to the control group. This has practical relevance for professional development, as increased use of on-the-fly strategies in transfer could imply that these strategies are more easily adapted and transferred by teachers than curriculum-embedded forms. Furthermore, it is important to note that our implementation quality scores assess a limited range of aspects which were judged central to our study. Regarding the assessment quality scale, this led to ceiling effects limiting the informative value of the results. In further analyses, it may be promising to investigate the quality of assessments and feedback in more detail, for example analysing the didactical quality of the assessment tasks and feedback.

Regarding the second research aim, the present study focused on the relationship of two selected teacher variables with implementation fidelity. In addition to MIS und EVA, other variables may be connected with teachers' implementation of formative assessment, including other aspects of teacher professional knowledge, beliefs and class characteristics. For example, students' level of achievement may have an impact on how teachers use and evaluate formative assessment.

As formative assessment was present in both units, albeit to a different extent, future analyses may now focus on the *achieved treatment* (Ruiz-Primo, 2006) – that is, did the formative assessment intervention positively influence students' outcomes? First results regarding direct application show that the formative assessment intervention presented here – sufficient implementation of at least 70% of intended elements provided – has positive effects on students' achievement as compared to the control group (Decristan et al., 2015). Further analyses should follow, including the transfer condition and investigating the relationship between implementation fidelity and student outcomes in more detail, to provide further insight into what may be considered a 'sufficient' implementation of formative assessment.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the Hessian initiative for the development of scientific and economic excellence (LOEWE).

Note

1. Due to organisational reasons, we could not observe the frequency of curriculum-embedded formative assessment in one control group class.

Notes on contributors

Annika Lena Hondrich received a Diploma in Psychology at Goethe University, Frankfurt am Main. Currently, she is a Ph. D. candidate at the German Institute for International Educational Research (DIPF) in Frankfurt am Main. Her research is centered on the implementation of formative assessment in classroom instruction and its impact on students' achievement and motivation.

Silke Hertel is a full professor for Personal Resources in School Context at Ruprecht Karls University Heidelberg. After receiving a Doctorate in Psychology at the Technische Universität Darmstadt, she was a post-doctoral fellow at the DIPF and assistant professor for Adaptive Learning Environments at the Center for Research on Individual Development and Adaptive Education of Children at Risk (IDeA), which was jointly established by DIPF and Goethe University. Her research projects focus on teacher professional development, teacher–parent collaboration, and adaptive learning environments.

Katja Adl-Amini is a Special Needs teacher with several years of teaching experience in inclusive classrooms. Currently, she is a Ph. D. candidate at the DIPF. Her research focuses on the implementation and impact of peer learning environments and the fostering of constructive social interactions among students.

Eckhard Klieme is the Director of the Department for Research on Educational Quality and Evaluation at the DIPF. He also is a full professor of Educational Science at Goethe University and a Co-founder of the IDeA Center. Klieme received a Diploma in Mathematics and a Doctorate in Psychology from the University of Bonn. He has been the principal investigator for national and international large scale surveys as well as for several video-based studies on classroom instruction. His current research is focused on teaching quality and school effectiveness.

References

- Anderson, R. D. (2002). Reforming science teaching: What research says about inquiry. *Journal of Science Teacher Education*, 13, 1–12. doi:10.1023/A:1015171124982
- Appleton, K. (2007). Elementary science teaching. In S. K. Abell & N. Lederman (Eds.), *Handbook of research on science education* (pp. 493–537). Mahwah, NJ: Lawrence Erlbaum Associates.
- Bell, B., & Cowie, B. (2001). The characteristics of formative assessment in science education. *Science Education*, 85, 536–553. doi:10.1002/sce.1022
- Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy & Practice*, 18, 5–25. doi:10.1080/0969594X.2010.513678
- Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & 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
- Coffey, J. E., Hammer, D., Levin, D. M., & Grant, T. (2011). The missing disciplinary substance of formative assessment. *Journal of Research in Science Teaching*, 48, 1109–1136. doi:10.1002/tea.20440
- Decristan, J., Hondrich, A. L., Büttner, G., Hertel, S., Klieme, E., Kunter, M., ... Hardy, I. (2015). Impact of additional guidance in science education on primary students' conceptual understanding. *The Journal of Educational Research*, Published online on March 11, 2015. doi: 10.1080/00220671.2014.899957
- Decristan, J., Klieme, E., Kunter, M., Hochweber, J., Büttner, G., Fauth, B., ... Hardy, I. (in press). Embedded formative assessment and classroom process quality: How do they

- interact in promoting students' science understanding? *American Educational Research Journal*.
- Desimone, L. (2009). Improving impact studies of teachers' professional development: Toward better conceptualizations and measures. *Educational Researcher*, 38, 181–199. doi:10.3102/0013189X08331140
- Dixon, R. H., Hawe, E., & Parr, J. (2011). Enacting assessment for learning: The beliefs practice nexus. *Assessment in Education: Principles, Policy & Practice*, 18, 365–379. doi:10.1080/0969594X.2010.526587
- Dusenbury, L., Brannigan, R., Falco, M., & Hansen, W. B. (2003). A review of research on fidelity of implementation: Implications for drug abuse prevention in school settings. *Health Education Research*, 18, 237–256. doi:10.1093/her/18.2.237
- Falk, A. (2011). Teachers learning from professional development in elementary science: Reciprocal relations between formative assessment and pedagogical content knowledge. *Science Education*, 96, 265–290. doi:10.1002/sce.20473
- Fauth, B., Decristan, J., Rieser, S., Klieme, E., & Büttner, G. (2014). Student ratings of teaching quality in primary school: Dimensions and prediction of student outcomes. *Learning and Instruction*, 29, 1–9. doi:10.1016/j.learninstruc.2013.07.001
- 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
- Grambsch, P. M. (1994). Simple robust tests for scale differences in paired data. *Biometrika*, 81, 359–372. doi:10.1093/biomet/81.2.359
- Gresham, F. M. (1989). Assessment of treatment integrity in school consultation and prereferal intervention. *School Psychology Review*, 18, 37–50.
- Gresham, F. M. (2009). Evolution of the treatment integrity concept: Current status and future directions. *School Psychology Review*, 38, 533–540.
- Hardy, I., Hertel, S., Kunter, M., Klieme, E., Warwas, J., Büttner, G., & Lühken, A. (2011). Adaptive Lerngelegenheiten in der Grundschule: Merkmale, methodisch-didaktische Schwerpunktsetzungen und erforderliche Lehrkompetenzen [Adaptive learning environments in primary school: Characteristics, didactical approaches and required teacher competencies.] *Zeitschrift für Pädagogik*, 57, 819–833.
- Hardy, I., Jonen, A., Möller, K., & Stern, E. (2006). Effects of instructional support within constructivist learning environments for elementary school students' understanding of "floating and sinking". *Journal of Educational Psychology*, 98, 307–326. doi:10.1037/0022-0663.98.2.307
- Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77, 81–112. doi:10.3102/003465430298487
- Helmke, A. (2006). Unterrichtsqualität: Erfassen, Bewerten, Verbessern [Teaching quality: Assessment, evaluation, improvement]. Seelze: Kallmeyersche Verlagsbuchhandlung.
- Jonen, A., & Möller, K. (2005). *Die KiNT-Boxen – Kinder lernen Naturwissenschaft und Technik* [The KiNT-Boxes – Children learn science and technology]. Essen: Spectra-Verlag.
- Keeley, P. (2005). *Uncovering student ideas in science, volume 1*. Arlington, VA: National Science Teacher Association (NSTA).
- 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.
- 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
- Morrison, J. A., & Lederman, N. G. (2003). Science teachers' diagnosis and understanding of students' preconceptions. *Science Education*, 87, 849–867. doi:10.1002/sce.10092

- Mowbray, C. T., Holter, M. C., Teague, G. B., & Bybee, D. (2003). Fidelity criteria: Development, measurement, and validation. *American Journal of Evaluation*, 24, 315–340. doi:10.1177/109821400302400303
- Noyce, P. E. (2011). Introduction and overview: The elusive promise of formative assessment. In P. E. Noyce & D. T. Hickey (Eds.), *New frontiers in formative assessment* (pp. 1–12). Cambridge, MA: Harvard Education Press.
- O'Donnell, C. (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
- Penuel, W. R., Fishman, B. J., Yamaguchi, R., & Gallagher, L. P. (2007). What makes professional development effective? Strategies that foster curriculum implementation. *American Educational Research Journal*, 44, 921–958. doi:10.3102/0002831207308221
- Postholm, M. B. (2012). Teachers' professional development: A theoretical review. *Educational Research*, 54, 405–429. doi:10.1080/00131881.2012.734725
- Ruiz-Primo, M. A. (2006). *A multi-method and multi-source approach for studying fidelity of implementation*. CSE: Technical Report 677. Los Angeles, CA: Center for Research on Evaluation, Standards, and Student Testing/University of California.
- Pryor, J., & Crossouard, B. (2008). A socio-cultural theorisation of formative assessment. *Oxford Review of Education*, 34, 1–20. doi:10.1080/03054980701476386
- Ruiz-Primo, M. A., Furtak, E. M., Ayala, C., Yin, Y., & Shavelson, R. J. (2010). Formative assessment, motivation, and science learning. In H. L. Andrade & G. J. Cizek (Eds.), *Handbook of formative assessment* (pp. 139–158). New York, NY: Routledge.
- Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, 18, 119–144. doi:10.1007/BF00117714
- 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
- Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher*, 15, 4–14. doi:10.3102/0013189X015002004
- Smith, E., & Gorard, S. (2005). 'They don't give us our marks': The role of formative feedback in student progress. *Assessment in Education: Principles, Policy & Practice*, 12, 21–38. doi:10.1080/0969594042000333896
- Tierney, R. D. (2006). Changing practices: Influences on classroom assessment. *Assessment in Education: Principles, Policy & Practice*, 13, 239–264. doi: 10.1080/09695940601035387
- Tomita, M. K. (2009). *Examining the influence of formative assessment on conceptual accumulation and conceptual change* (Doctoral dissertation). Retrieved from ProQuest Dissertation and Theses database on <http://pqdtopen.proquest.com/pubnum/3343949.html>
- Torrance, H., & Pryor, J. (2001). Developing formative assessment in the classroom: Using action research to explore and modify theory. *British Educational Research Journal*, 27, 615–631. doi:10.1080/01411920120095780
- Vosniadou, S. (Ed.). (2008). *International handbook of research on conceptual change*. New York, NY: Routledge.
- Watters, J. J., & Ginns, I. S. (1997). An in-depth study of a teacher engaged in an innovative primary science trial professional development project. *Research in Science Education*, 27, 51–69. doi:10.1007/BF02463032
- William, D., Lee, C., Harrison, C., & Black, P. (2004). Teachers developing assessment for learning: Impact on student achievement. *Assessment in Education: Principles, Policy & Practice*, 11, 49–65. doi:10.1080/0969594042000208994
- Yin, Y., Shavelson, R. J., Ayala, C. C., Ruiz-Primo, M. A., Tomita, M., Furtak, E. M., ... Young, D. B. (2008). On the measurement and impact of formative assessment on students' motivation, achievement, and conceptual change. *Applied Measurement in Education*, 21, 335–359. doi:10.1080/08957340802347845

Appendix 1.

Table A1. Observation items used for measuring implementation fidelity in the treatment group.

Score	Items	Source	Intended treatment maximum score
Frequency – assessment	Assessment on content of lesson 1 Assessment on content of lesson 2 Assessment on content of lesson 3 Assessment on content of lesson 4	Document analysis (students' workbooks)	4 = 100%
Frequency – feedback	Written, individual feedback on assessment 1 ^a Written, individual feedback on assessment 2 Written, individual feedback on assessment 3 Written, individual feedback on assessment 4	Document analysis	3 = 100%
Frequency – adaptive instruction	Assignment of differentiated worksheets according to assessment 1 ^a Assignment of differentiated worksheets according to assessment 2 Assignment of differentiated worksheets according to assessment 3 Assignment of differentiated worksheets according to assessment 4	Document analysis	3 = 100%
Quality of assessment	Covers target concept of the respective lesson Includes tasks suited for assessing relevant students' misconceptions	Document analysis	2
Quality of feedback	Includes knowledge of response Includes information on students' level of understanding Includes hint or strategy for improvement	Document analysis	3
Quality of enactment - transparency	Teacher discusses assessments/feedback with students Teacher explicitly states the formative purpose of assessment	Classroom observation	2

^aAs the first assessment was intended to evaluate students' preconceptions rather than the understanding of a specific, scientifically desired concept and was followed by experiments rather than worksheets, no feedback and assignment of differentiated tasks were expected. Therefore, a feedback and adaptive instruction score of 3 reflects 100% of the intended treatment. However, as teachers were free to provide feedback and differentiated tasks also after assessment 1, they were nevertheless included as items.

Brain teaser No. 4



1. Which cubes float?

a) Here are some more cubes. They are just as big as our water cube. Check the correct box: do the cubes float or sink in water?

water cube



61g

brown plastic cube



49g

- floats
 sinks

red plastic cube



62g

- floats
 sinks

clay cube



93g

- floats
 sinks

b) Here are cubes of different sizes. They each weigh 61 grams. Check for each cube - does it float or sink?

water cube



61g



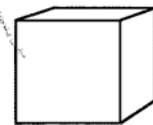
61g

- floats
 sinks



61g

- floats
 sinks



61g

- floats
 sinks

2. Why do some things float in water and others sink?

Write down your explanation.

Evaluation of Brain teaser No. 4: Relative Density

Task 1a repeats the simple application of the concept of relative density. It does not require relevant transfer. Task 1b, by contrast, requires a deeper understanding and the ability to transfer the concept.




Task 2 offers the opportunity for students to write down their own explanations of the floating and sinking of objects. It enables a qualitative evaluation of conceptions the students are currently using. A direct comparison with students' explanations in the first lesson is possible.

1. Evaluation of (pre)conceptions in Task 2:

Please note in the learning progression table, which (pre-)conceptions students argue with. For examples on preconceptions: see Evaluation of Brain teaser 1 (Lesson 1).

Conceptions	Students' answers (examples)
Material concept	"Things float because they are made of wood , and sink because they are made of metal ."
Simple density	"Things float if they are light for their size ."
Relative density	"Things float if they are lighter than the same amount of water . They sink if they are heavier than water ."

2. Evaluation of the understanding of relative density:

- **No mistake in task 1a and 1b:** the concept of relative density is well understood and can be applied flexibly.
 - In the table: please check the field for relative density
 - Assign worksheet tasks for further application and transfer to new situations (worksheets with a square). 
 - Provide feedback: template 2A
- **Task 1a correct, one or more mistakes in task 1b:** the concept of relative density has been understood, but cannot yet be applied to tasks requiring transfer.
 - In the table: please enter a check in parentheses in the field for relative density
 - Tasks fostering the confident application of the concept (worksheets with the triangle). 
 - Provide feedback: template 2B
- **One or more mistakes in task 1a:** the concept of relative density is not entirely understood and cannot yet be applied.
 - In the table: please enter \emptyset in the field for relative density
 - Assign repetition and simple application tasks (worksheets with the circle). 
 - Provide feedback: template 2C

If a student's level of understanding is unclear because he or she shows an inconsistent pattern of answers (e.g., task 1b correct, task 1a wrong), it is helpful to additionally consider the open answer of task 2. If in doubt, simple application tasks should be assigned to make sure that the student learns how to apply the concept.