Decristan, Jasmin; Klieme, Eckhard; Kunter, Mareike; ...

Embedded formative assessment and classroom process quality. How do they interact in promoting students' science understanding

American educational research journal 52 (2015) 6, S. 1-27

American Educational Research Journal
Month XXXX, Vol. XX, No. X, pp. 1–27
DOI: 10.3102/0002831215596412

Embedded Formative Assessment and Classroom Process Quality: How Do They Interact in Promoting Science Understanding?

Jasmin Decristan
German Institute for International Educational Research
Eckhard Klieme
German Institute for International Educational Research
Institute of Educational Sciences, Goethe University Frankfurt
Mareike Kunter
Institute of Psychology, Goethe University Frankfurt
Jan Hochweber
Institute of Research on Teaching Profession and on Development of Competencies, University of Teacher Education St. Gallen
Gerhard Büttner
Benjamin Fauth
Institute of Psychology, Goethe University Frankfurt
A. Lena Hondrich
German Institute for International Educational Research
Svenja Rieser
Institute of Psychology, Goethe University Frankfurt
Silke Hertel
Institute of Educational Science, University of Heidelberg
Ilonca Hardy
Institute of Educational Sciences, Goethe University Frankfurt

In this study we examine the interplay between curriculum-embedded formative assessment—a well-known teaching practice—and general features of classroom process quality (i.e., cognitive activation, supportive climate, classroom management) and their combined effect on elementary school students' understanding of the scientific concepts of floating and sinking. We used data from a cluster-randomized controlled trial and compared curriculum-embedded formative assessment (17 classes) with a control group (11 classes). Curriculum-embedded formative assessment and classroom process quality promoted students'

learning. Moreover, classroom process quality and embedded formative assessment interacted in promoting student learning. To ensure effective instruction and consequently satisfactory learning outcomes, teachers need to combine specific teaching practices with high classroom process quality.

Jasmin Decristan is senior researcher at the German Institute for International Educational Research (DIPF) and a member of the Center for Individual Development and Adaptive Education of Children at Risk (IDeA), Schloßstr. 29, 60486 Frankfurt a.M., Germany; e-mail: decristan@dipf.de. *Her research focuses on individual support of students, differential effects of instruction, teaching quality, and peer-assisted learning.*

Eckhard Klieme is director of the Department of Educational Quality and Evaluation at the DIPF, a member of the scientific board of IDeA in Frankfurt, Germany, and professor of educational science at the Goethe University in Frankfurt, Germany. His research focuses on teaching quality, school effectiveness, school development, and international comparative education research.

Mareike Kunter is professor of educational psychology at the Goethe University and a member of the scientific board of IDeA in Frankfurt, Germany. Her research focuses on professional competence and professional development of teachers and research on learning and instruction in the school context.

Jan Hochweber is head of the Department Learning and Assessment Systems at the University of Teacher Education in St. Gallen, Switzerland. His research focuses on classroom assessment, teaching quality, school effectiveness, and quantitative methods in education research.

Gerhard Büttner is professor of educational psychology at the Goethe University and a member of IDeA in Frankfurt, Germany, and director at the Academy of Education Research and Teacher Education. His research focuses on self-regulated learning, teaching quality, working memory, learning disabilities, and intellectual disabilities.

Benjamin Fauth is now assistant professor at the Hector Research Institute of Education Sciences and Psychology at the University of Tübingen and a member of IDeA in Frankfurt, Germany. His research focuses on teaching quality and self-regulated learning in school.

A. Lena Hondrich is PhD student at the DIPF and a member of IDeA in Frankfurt, Germany. Her research focuses on embedded formative assessment, motivation, and science education.

Svenja Rieser is now a member of the research staff at the TU Dortmund University in Dortmund and a member of IDeA in Frankfurt, Germany. Her research focuses on teaching quality in school and preschool, self-regulated learning, and metacognition in school learning.

Silke Hertel is professor for personal competencies in school context at the Heidelberg University and a member of IDeA in Frankfurt, Germany. Her research focuses on professional competencies and professional development of teachers, cooperation between families and schools, and the arrangement of (adaptive) learning environments.

Ilonca Hardy is professor of elementary education at the Goethe University and a member of IDeA in Frankfurt, Germany. Her research focuses on instructional design in preschool and elementary school, science education, and bilingual education.

KEYWORDS: assessment, classroom research, science education, school/teacher effectiveness

From both a teaching effectiveness point of view (e.g., Brophy, 2000; Scheerens & Bosker, 1997; Wang, Haertel, & Walberg, 1993) and a teacher education point of view (Ball & Forzani, 2011), effective teaching can be described as practices predictive of student learning, which can be developed in teacher training or professional development programs (Grossman, Loeb, Cohen, & Wyckoff, 2013). However, approaches currently taken to describe effective teaching differ in scope and depth. While for instance Ball and Forzani (2011) focus on specific "high-leverage practices" such as choosing and representing content, organizing small-group work, or employing certain methods to assess students, Pianta, La Paro, and Hamre (2008) provide a rating scheme according to broad, global dimensions, namely classroom organization, emotional support, and instructional support. The Danielson Framework (1996), which greatly influenced the Measures of Effective Teaching project (Kane, McCaffrey, Miller, & Staiger, 2013), is a mix of specific teaching practices and global factors. In a synthesis of research on effective teaching, Good, Wiley, and Florez (2009) noted nine general principles, again including specific teaching practices (i.e., scaffolding students' ideas and task involvement, practice/application, goal-oriented assessments) and global factors of classroom process quality (i.e., thoughtful discourse, proactive and supportive classrooms, classroom management). The authors also included content-related principles (i.e., coherent content, curriculum alignment, appropriate expectations). Particularly classroom process quality and specific teaching practices have received much attention in empirical research (Baumert et al., 2010; Grossman et al., 2013; Hattie, 2009; Reyes, Brackett, Rivers, White, & Salovey, 2012).

Despite the large body of research in this area, the interplay between global factors of effective teaching and specific teaching practices in enhancing students' learning rarely has been investigated. In this article we disentangle both categories within a quasi-experimental study of early science education. Teachers obtained training on curriculum-embedded formative assessment, a prominent example of a specific teaching practice, and implemented it in their science classes. The general dimensions of classroom process quality were rated by students. To control content matter, all the teachers conducted a scripted unit on floating and sinking, which has been shown to enhance students' understanding of these science concepts (Hardy, Jonen, Möller, & Stern, 2006; for American-based research on this topic, see Shavelson et al., 2008; Shemwell & Furtak, 2010).

Curriculum-Embedded Formative Assessment as a Specific Teaching Practice

Since the seminal work of Black and Wiliam (1998), formative assessment has become one of the most prominent teaching practices in education research. It is defined as the repeated use of assessment-based information to recognize and respond to students' needs to enhance learning (see Bell & Cowie, 2001, p. 536). A meta-analysis has shown that formative assessments supports student learning (Kingston & Nash, 2011); however, the term *formative assessment* has been used inconsistently in the literature, and this approach has been operationalized in diverse ways (Kingston & Nash, 2011). Classroom formative assessment can be distinguished by its degree of formality ranging from informal "on the fly" assessment during classroom instruction and discourse to "curriculum-embedded" assessment as formal and planned diagnostic tests placed at specific joints in the curriculum where a central subgoal of learning should have been met (e.g., Shavelson et al., 2008; Wilson & Sloane, 2000).

Embedded formative assessments can be implemented in the curriculum in several ways described in terms of structural and quality components (e.g., Furtak et al., 2008). Structural components of embedded formative assessments refer, for instance, to the time frame (between lessons, between teaching units, or over semesters/years), frequency, and methods of formal assessment and feedback (e.g., paper-pencil test, computer-based assessments, and teacher- or peer-mediated feedback). Attempts are currently being made to understand why and how formative assessment affects students' learning processes (e.g., Black & Wiliam, 2009; Rakoczy, Harks, Klieme, Blum, & Hochweber, 2013). Black and Wiliam (2009) present several key strategies that refer to the quality of implementing formative assessment in a curriculum. These key strategies aim to make students' current level of understanding and their learning processes more explicit, to articulate clearly learning goals to students, and to engage them in learning (see also Furtak et al., 2008). In particular, addressing the key strategies of formative assessment when providing feedback has a great impact on student learning (e.g., Hattie, 2009; Hattie & Timperley, 2007). To be effective, feedback on assessment should be given in a timely manner; it informs students about their current conceptions and competencies, their learning progress, and the learning goals, and assists and encourages students to take the next learning step (Hattie & Timperley, 2007; Sadler, 1989). Such feedback strategies have been successfully implemented in formative assessment (e.g., Rakoczy et al., 2013).

Global Dimensions of Classroom Process Quality

Recent international research has identified key features or basic dimensions of effective teaching, which focus on classroom process quality and

build on the assumptions of cognitive-constructivist or socio-constructivist models of teaching and learning (e.g., Bransford, Brown, & Cocking, 2000). Several authors have identified independently but consistently three global dimensions of classroom process quality in Europe (i.e., cognitive activation, supportive climate, and classroom management; Klieme, Pauli, & Reusser, 2009; for a similar model, see Baumert et al., 2010) and in the United States (i.e., instructional support, emotional support, and classroom organization (Pianta et al., 2008; Reyes et al., 2012). These processes are described below.

Cognitive Activation or Instructional Support

Cognitive activation refers to instructional strategies to develop students' conceptual understanding. Teachers explore and build on students' prior concepts and ideas, specify connections among facts and procedures, compare ideas and concepts, and use challenging tasks and nonroutine problems to stimulate cognitive conflicts and engage students in higher-level thinking processes (Baumert et al., 2010; Lipowsky et al., 2009). Through participation in classroom discussion, students may communicate concepts and ideas and develop conceptual understanding (e.g., Osborne, Erduran, & Simon, 2004, in the context of science education). Empirical findings have confirmed that cognitive activation fosters student learning (e.g., Baumert et al., 2010; Kunter et al., 2013; Lipowsky et al., 2009; Pianta & Hamre, 2009). Many of these studies have been conducted in secondary school mathematics classrooms. However, the concept of cognitive activation can be successfully applied to other subjects in elementary school (see Fauth, Decristan, Rieser, Klieme, & Büttner, 2014; Hamre, Pianta, Mashburn, & Downer, 2007; Reyes et al., 2012).

Supportive Climate or Emotional Support

Teacher-learner interactions have been conceptualized and theoretically framed in markedly different ways (see Cornelius-White, 2007; Davis, 2003). A supportive classroom climate requires positive teacher-student relationships and constructive learner support. It is characterized by a warm classroom atmosphere and caring teacher behavior, constructive teacher feedback, and a positive approach to student errors and misconceptions (e.g., Brophy, 2000; Klieme et al., 2009; Lipowsky et al., 2009). Studies on the impact of a supportive climate confirm its positive effect on student learning outcomes (e.g., Cornelius-White, 2007; Pianta, Nimetz, & Bennett, 1997; Reyes et al., 2012).

Classroom Management or Classroom Organization

Effective classroom management has been acknowledged for several decades as a central feature of successful instruction (e.g., Emmer & Stough, 2001; Hattie, 2009; Kounin, 1970). This dimension of classroom

process quality is closely connected with the concept of time on task (Seidel & Shavelson, 2007; Wang et al., 1993). Classroom management strategies involve the implementation of clear rules and procedures in the classroom and the use of smooth transitions between activities (Doyle, 1986). Furthermore, effective classroom management requires strategies for coping with disruptions and dealing with disciplinary problems (Emmer & Stough, 2001; Kounin, 1970). Effective classroom management has been shown to support student learning at different school levels and in various domains (e.g., Hattie, 2009; Seidel & Shavelson, 2007; Wang et al., 1993).

Interplay Between Classroom Process Quality and Embedded Formative Assessment

To gain further insight into how teaching affects student learning, Raudenbush (2008) suggested that researchers focus on *enacted* regimes (i.e., regimes that students actually experience) rather than on *intended* regimes (i.e., planned treatment or curricula). Regarding the enactment of specific teaching practices, Furtak et al. (2008) highlighted the importance of quality features (the quality of enactment) of embedded formative assessment rather than structural components in supporting students' science understanding. However, determining the quality of enactment requires appropriate measures as well as elaborate and time-consuming analyses of classroom instruction (Raudenbush, 2008). This holds particularly true for assessing the quality of specific teaching practices with instructionally sensitive, reliable, and valid instruments to be tailored each time they are employed. Conceptually, the characteristics of a high-quality enactment of embedded formative assessment may be closely related to global factors characterizing high classroom process quality, particularly cognitive activation and supportive climate. To activate students cognitively, teachers must explore and build on students' prior understanding and engage them in higher-level thinking processes (Brophy, 2000; Lipowsky et al., 2009). Likewise, the diagnostic tasks and feedback in embedded formative assessment aim to activate students cognitively by making the learning processes more explicit to students and by engaging them in learning (Black & Wiliam, 2009). Furthermore, a supportive climate is characterized by a positive teacher-student relationship as well as constructive learner support through constructive teacher feedback and a positive approach to student errors and misconceptions (e.g., Brophy, 2000; Davis, 2003; Klieme et al., 2009; Lipowsky et al., 2009). Similarly, key strategies of embedded formative assessment are used for constructive feedback and learner support (Black & Wiliam, 2009; Hattie & Timperley, 2007). In contrast, the quality of enactment of embedded formative assessment cannot be clearly assigned to efficient classroom management.

In general, according to meta-analytic findings, quality processes of classroom instruction are one of the most powerful predictors of student learning with medium to large effect sizes (e.g., Hattie, 2009; Seidel & Shavelson, 2007). Moreover, classroom processes moderate the relationship between student characteristics and learning outcomes: Students at risk of failure at school show greater achievement in high-quality classrooms than their peers in low-quality classrooms (e.g., Curby, Rimm-Kaufman, & Ponitz, 2009; Hamre & Pianta, 2005). More importantly, classroom processes moderate the effectiveness of "regimes" (Raudenbush, 2008; e.g., treatments, specific teaching practices, or curricula). O'Donnell (2007) showed that high-quality instructional strategies moderated the relationship between a science curriculum condition and middle school students' achievement. In their meta-analysis of clinical treatments, Landenberger and Lipsey (2005) identified the quality of enactment ("higher quality implementation"; p. 469) as one of the most powerful moderators of treatment effectiveness. The same principle also should apply to the implementation of specific treatments such as embedded formative assessment in educational practice.

Aim of This Study and Hypotheses

The aim of this study is to examine the interplay between global factors of classroom process quality (i.e., an established three-dimensional model; e.g., Klieme et al., 2009; Pianta et al., 2008) and the specific teaching practice of embedded formative assessment (e.g., Black & Wiliam, 1998; Kingston & Nash, 2011) in promoting students' understanding of the scientific concepts of floating and sinking. We support calls for high-quality (quasi-)experimental studies (e.g., Kingston & Nash, 2011; Raudenbush, 2008) to draw more valid conclusions. We therefore conducted a cluster-randomized intervention study in elementary science classes with standardized treatments and a control group to test the following hypotheses:

Hypothesis 1: The specific teaching practice of embedded formative assessment will be effective in supporting students' science understanding.

Hypothesis 2: Global dimensions of classroom process quality (i.e., cognitive activation, supportive climate, and effective classroom management) will positively predict students' science understanding.

Hypothesis 3: The effectiveness of embedded formative assessment will be moderated positively by classroom process quality (i.e., cognitive activation and supportive climate). Specifically, the effectiveness of embedded formative assessment on students' science understanding, when compared with the control group, will increase with higher levels of cognitive activation or supportive climate.

Hypothesis 1 was confirmed in a previous publication focusing on the effects of various teaching practices on student learning in a subsample of

the intervention study. In this previous publication, only those classes (n = 25) were considered that showed a minimum implementation of 70% of the treatment and the content matter when compared with the scripted unit on floating and sinking (see Decristan et al., 2015). Still, it is worth exploring this hypothesis again to understand better the underlying processes involved in embedded formative assessment and how it enhances student learning.

Method

Study Design

This study was part of an intervention study that used a cluster-randomized controlled trial to compare the effects of different teaching practices on elementary school students' conceptual understanding of the floating and sinking of objects (Project IGEL; Individual support and adaptive learning environments in primary school). The target population of the study was students in public elementary schools in a federal state of Germany. Teachers and principals were contacted by telephone and invited to attend information sessions on the research project. Each school that volunteered to participate in the study was randomly assigned to one of the three instructional conditions of the intervention or to the control group. In the present article, we focus on one instructional condition, that is, embedded formative assessment, and the control group. All teachers, including those of the control group, participated in professional development workshops to learn about the particular intervention condition. Subsequently, they conducted the predetermined ready-made unit in their elementary science classes, following instructions set out in a manual and implementing preselected teaching material. Student data were assessed before and after the unit.

Participants

The sample participating in the present study consisted of 28 teachers and 551 third grade students from 18 public elementary schools in a federal state of Germany (embedded formative assessment: 17 teachers, 319 students; control group: 11 teachers, 232 students).[1] The mean class size was 20 students (class size ranged from 10 to 26 students). All of the schools were located in central Germany in rural (57% of classes) and urban areas. Participation was voluntary for teachers and students. The mean age of the teachers (86% female) was 43.4 years (SD = 9.8), and their mean professional experience was 15.8 years (SD = 9.8). All of the teachers had been teaching science for the previous five years. The students (48% female) were 8.8 years old on average (SD = 0.5) and 38% (formative assessment group: 44%, control group: 30%) reported that at least one of their parents had been born outside of Germany, and 23% (formative assessment group: 28%, control group: 15%) reported that both of their parents had been born

outside of Germany. The sample included students from diverse ethnic backgrounds, mostly from immigrant families speaking Turkish or a Romanic, Slavic, African, or other (Indo-)Germanic language.

Professional Development Workshops for Teachers

All teachers participating in the study attended a professional development workshop on floating and sinking. They also received training on embedded formative assessment or control group content (parental counseling). Workshops were held over 4 days (4.5 hours per day) by experienced trainers. On the first day subject matter from the curriculum, that is, the concept of density (material, density, and density of water relative to the density of objects), was covered. On the second day embedded formative assessment or control group content was covered. On the third and fourth days the use of embedded formative assessment for classroom instruction on floating and sinking and the control group content were covered.

Unit on Floating and Sinking

The unit, which was implemented in both the experimental group and the control group to control for content matter and sequencing of content, was designed according to key principles of inquiry-based science education (for more details, see Decristan et al., 2015; Hondrich, Hertel, Adl-Amini, & Klieme, in press). The unit was adapted from an empirically evaluated elementary school unit on the floating and sinking of objects (Hardy et al., 2006) and covered conceptual aspects of inquiry. The unit included one introductory lesson of 45 minutes and four subsequent lessons of 90 minutes. To standardize the unit, all teachers were provided with a detailed manual that included prestructured lessons with detailed lesson plans and the learning goals of the four lessons, worksheets on floating and sinking at three different levels of conceptual understanding, and a box of objects made of different materials (e.g., steel, Styrofoam, and wax; see Möller & Jonen, 2005) for teacher demonstrations and hands-on activities for the students.

Curriculum-Embedded Formative Assessment Group and Control Group

After the workshop on the unit on floating and sinking, the teachers participated in a training session on embedded formative assessment or control group content.

Embedded formative assessment. Teachers in the experimental group received information on the design and formative use of diagnostic tasks in the curriculum. The open-ended assessments were designed to elicit students' current conceptions and to examine their current level of conceptual understanding. Furthermore, teachers were provided with information on formal strategies for supplying feedback to students and on adaptation of

subsequent instruction. Using a semistructured feedback sheet, teachers gave students written feedback on their assessment results and input on their subsequent learning steps, and they assigned differentiated tasks that matched the students' current level of conceptual understanding of floating and sinking (see Hondrich et al., in press, for details). The teachers were instructed to embed four formal assessments at specific joints in the unit.

Control group. The control group teachers were instructed to use the standardized science unit but were not given any instruction on using specific teaching practices. Instead, they completed a workshop on parental counseling (see Hertel, 2009), which was not expected to affect students' science understanding. We thus aimed to keep workshop times comparable between the intervention groups as well as to provide content that was related to teachers' everyday professional life.

All of the teachers conducted the science unit (with or without embedded formative assessment) in their classes in the second term of the academic year over a span of 4 weeks at most.

Manipulation Checks

Each class was either video recorded or visited by project members during one of the 90-minute units.[2] A checklist with a standardized list of dichotomous items (0 = did not occur, 1 = occurred) was used to check for adherence to the intended science lesson and the embedded formative assessment treatment. At least 45% of the classes were scored by two independent raters. Interrater agreement was higher than 85% for each item. For each class, a percentage score was computed for the lesson content (see Hondrich et al., in press). Teachers' adherence to the intended science lesson was $M = 86.15$ ($SD = 15.53$, min = 25, max = 100, $N = 28$). The Kolmogorov-Smirnov Z test demonstrated that both groups had similar content scores ($Z = .88$, $p = .415$). The embedded formative assessment checklist included four dichotomous items addressing the occurrence or nonoccurrence of two key categories of diagnosis and formative use of diagnostic information (see Hondrich et al., in press). A percentage score was computed for each class, and results demonstrated that the treatment components were implemented in the embedded formative assessment classes during the observed lesson ($M = 95.59$, $SD = 13.21$, min = 50, max = 100, $N = 17$). As expected, embedded formative assessment was not used in the control group classes ($M = 0$, $SD = 0$).

Instruments

Classroom Process Quality

Upon completion of the science unit, students rated the classroom process quality on a questionnaire that covered three subscales: cognitive

activation (seven items, Cronbach's α = .79), supportive climate (nine items, Cronbach's α = .89), and classroom management (five items, Cronbach's α = .88; see Fauth et al., 2014). Cognitive activation items referred to the teachers' exploration of students' prior concepts and ideas as well as presentation of challenging tasks. Supportive climate items were related to warm and caring teacher behavior, constructive feedback, and learner support. Classroom management items were related to the lack of disciplinary problems and disruptions during classroom instruction. Students were instructed to focus on the specific science unit on floating and sinking. All items were rated on a 4-point scale with categories ranging from 1 (*strongly disagree*) to 4 (*strongly agree*). The intraclass correlations indicated a substantial amount of variance among the classes (ICC1) and good reliability of the aggregated ratings on the classroom level (ICC2; see Lüdtke, Trautwein, Kunter, & Baumert, 2006) for all three scales: cognitive activation (ICC1 = .12, ICC2 = .73), supportive climate (ICC1 = .18, ICC2 = .80), and classroom management (ICC1 = .31, ICC2 = .89).

Student Tests

To measure students' learning outcomes, a *test of science understanding* was adapted from previously published and current research on the topic of floating and sinking of objects (Hardy et al., 2006; Schneider & Hardy, 2013). The 13 test items included multiple-choice items and two open-ended tasks. Experts from educational practice and research in science education had judged the items as valid and highly relevant for the topic of floating and sinking. Furthermore, in a validation study, Pollmeier et al. (in press) showed that responses to paper-pencil items were mostly in line with students' responses during interviews to assess conceptual understanding. Hardy et al. (2006) showed that the test of science understanding of floating and sinking was related to a transfer test focusing on students' application of the science concepts in a wider context. The items on the test of science understanding were based on a model of three different levels of conceptual understanding of floating and sinking with students' responses to the items scored as either naïve conceptions (0), everyday life conceptions (1), or scientific conceptions (2). In our study, two independent raters coded the open-ended tasks (kappa = .87; see Gwet, 2012) according to the three levels of conceptual understanding. Test items were scaled using the Partial Credit Model (Masters, 1982), and a weighted likelihood estimate (WLE; Warm, 1989) was computed for each student (EAP/PV reliability = .70). The relationship between the WLE scores and students' grades in science ($|r| = .46$) was found to be comparable with findings from previous studies on science competencies (e.g., Schütte, Frenzel, Asseburg, & Pekrun, 2007).

We considered as covariates students' cognitive ability, science competence, and language proficiency, which are known to be strongly associated

with science understanding. Research has shown that general cognitive abilities are strongly connected to abilities in specific subjects at school, especially mathematics, science, and language (e.g., Gustafsson & Balke, 1993). Cognitive abilities were assessed using the CFT-20R diagnostic test (Weiß, 2006), a German version of the Culture Fair Intelligence Test, which includes 56 items (Cronbach's $\alpha = .72$). The *science competency test* was adapted from TIMSS 2007 (Martin, Mullis, & Foy, 2008) comprising the cognitive domains of knowing, applying, and reasoning. The test is in line with the science curriculum in elementary school in Germany. Experts from educational practice and research in science education had judged the test items used in our study to be highly relevant for elementary school science education and appropriate for third grade students. All items were piloted in class-wide assessments at the end of second grade. The test was composed of 12 items that fit the 1PL-Rasch model, and a WLE was computed for each student (EAP/PV reliability = .70). Furthermore, science education requires being able to analyze, summarize, and present information in oral or written formats (Lee, 2005) and thus is closely connected with language proficiency (e.g., Martin, Mullis, Foy, & Stanco, 2012). The language proficiency test was multiple choice and assessed students' passive vocabulary and sentence comprehension in German. It was adapted from diagnostic tests of German language comprehension (Elben & Lohaus, 2001; Glück, 2011; Petermann, Metz, & Fröhlich, 2010). For each of the 20 items, a verbal stimulus (a word or sentence) and a set of four pictures were presented. The students were instructed to choose the picture that matched the verbal stimulus. Answers were coded dichotomously (0 = not correct, 1 = correct). A total score was computed for each student (Cronbach's $\alpha = .72$). The relationship between this total score and students' family background (a dichotomous variable; 0 = both parents born in Germany, 1 = at least one parent born outside of Germany) was $|r| = .39$. Although we randomly assigned participating schools to the intervention conditions, both groups differed in students' test scores prior to the intervention: Students in the embedded formative assessment group scored higher than those in the control group on the test of cognitive ability, $t(508) = -3.63$, $p < .01$, and significantly lower on the tests of science competence, $t(516) = 2.51$, $p < .05$, and language proficiency, $t(515) = 2.50$, $p < .05$.

Procedure

The intervention took place during the academic school year 2010–2011. Data were collected from students by trained research staff members following standardized instructions. The items were read aloud by staff members and were presented visually with a projector. Each assessment required approximately 90 minutes to complete. The individual-level control variables were assessed at the beginning of the academic school year

(September–October 2010). In January and February 2011, the teachers conducted once or twice a week over a maximum span of 4 weeks the science unit comprising 4.5 lessons of 90 minutes each. Conceptual understanding of floating and sinking was assessed upon completion of the unit (February–March 2011).

Data Analyses

To account for the hierarchical data structure, we applied multilevel regression analysis (Raudenbush & Bryk, 2002) with students (Level 1) nested in classes (Level 2) and students' science understanding as the outcome variable. All models were estimated in Mplus 7.11 (Muthén & Muthén, 1998–2013). We were interested mainly in the relationship between classroom-level variables (classroom process quality and treatment) and science understanding, controlling for the individual-level covariates. The treatment variable was dummy coded (0 = control group, 1 = embedded formative assessment). All of the individual-level variables were standardized (M = 0, SD = 1) and centered at the grand mean (i.e., the mean score of the sample was subtracted from each individual score in a class; see Enders & Tofighi, 2007). The individual ratings of classroom process quality were aggregated to the classroom level to examine differences in ratings among classes (Lüdtke, Robitzsch, Trautwein, & Kunter, 2009). The aggregated ratings were then standardized (M = 0, SD = 1) and centered at the grand mean (i.e., the average of the class means was subtracted from each class mean).

First, we included the treatment variable (Hypothesis 1, Model 1) and the aggregated student ratings of each of the three dimensions of classroom process quality as independent variables in separate analyses (Hypothesis 2, Models 2 to 4). Next, we extended the models by simultaneously adding the treatment variable, one dimension of classroom process quality, and a second-level interaction variable that was the product of both classroom-level variables (treatment × classroom process quality; Hypothesis 3, Models 1 to 3).

A lack of test power is one of the major concerns of small sample sizes. Because of the small sample size at the classroom level, we conducted a power analysis using the Monte Carlo simulation method implemented in Mplus 7 (for an example, see Bolger, Stadler, & Laurenceau, 2012). Data simulations using 28 classes with an average class size of 20 students achieve a test power β of .60 to detect a medium to large effect of a class-level variable (R^2 = .215), and a power β of .77 to detect a large effect of a class-level variable (R^2 = .300).

As measure of effect size, we reported R^2. With respect to R^2, the explained variance in students' science understanding at each level of analysis was reported. In Mplus, the explained variance at the class level referred

to the explained variation in the random intercept after controlling for individual-level predictors.

Missing Data

The student participation rate was high (97% of the students in all 28 classes, the average participation rate was 96% in the embedded formative assessment group and 99% in the control group). The participation rate for each assessment point was at least 92% (min. 91% for each assessment point by treatment). Of the data, 2% were missing due to students changing schools or classes. For the variables used in our study, there were no missing data at the classroom level (aggregated student ratings of classroom process quality and the treatment variable). The amount of missing data at the individual level (i.e., students' test scores as covariates) ranged from 6.0% for students' science competence to 7.4% for cognitive ability. To deal with missing data at the individual level, we used the multiple implementation procedure in Mplus 7.11 to replace each missing value with a set of 10 predicted values. To this end, we specified an unrestricted ($H1$) model and included all variables used in our analysis as well as further individual-level auxiliary variables (Collins, Schafer, & Kam, 2001) in the imputation model.

Results

Descriptive Results

Table 1 presents the descriptive data analysis of both individual- and classroom-level variables. To provide a better interpretation of the descriptive data, the mean scores and standard deviations refer to the original metric of the variables. The students showed cognitive abilities that generally were comparable with the population of same-aged students in Germany (the diagnostic test is standardized to $M = 100$ and $SD = 15$; see Weiß, 2006). Regarding science competence, students answered on average 6 of 12 items, meaning they were able to solve tasks requiring mainly science knowledge. Students' mean score of science understanding (8 of 19 items) can be interpreted according to the model of different levels of conceptual understanding. This means that students on average rejected naïve conceptions but were inconsistent in the use of explanations of everyday life. Students' mean language proficiency (14.8 of a maximum score of 20) was rather high, but still showed substantial variation in individual scores. Students' ratings of cognitive activation and supportive climate also were rather high but at an expected level for elementary school student ratings (e.g., Doll, Spies, LeClair, Kurien, & Foley, 2010).

All of the variables exhibited substantial variance among classes, as indicated by the intraclass correlations (ICCs). In general, correlations between variables at the classroom level were greater than at the individual level. This

Table 1

Descriptive Data Analysis of Individual-Level Covariates and Perceived Classroom Process Quality

	1	2	3	4	5	6	7	*M*	*SD*	ICC
1. Cognitive ability	—	.56**	.47*	.67**	−.23	−.01	.50**	103.77	15.32	.15
2. Science competence	.43**	—	.86**	.56**	−.51**	−.40*	−.01	6.43	2.38	.16
3. Language proficiency	.40**	.61**	—	.40*	−.42*	−.33	.05	14.78	2.98	.20
4. Science understanding	.36**	.41**	.34**	—	−.05	.08	.35	8.00	3.75	.16
5. Cognitive activation	−.05	−.21**	−.19**	−.14**	—	.89**	.30	3.17	0.66	.11
6. Supportive climate	.05	−.17**	−.15**	−.02	.73**	—	.46*	3.28	0.67	.17
7. Classroom management	.04	−.08	−.02	.04	.36**	.44**	—	2.55	0.87	.31

Note. Correlations between individual-level variables are listed below the diagonal; correlations between class-level aggregated variables are listed above the diagonal. Means and standard deviations refer to the individual-level variables. In order to provide a better interpretation of the descriptive data, means and standard deviations refer to the original metric of the variables. The intraclass correlations (ICCs) indicate the proportion of variance among classes.

*$p < .05$. **$p < .01$.

holds particularly true for the three dimensions of classroom process quality. However, Fauth et al. (2014) have shown with a larger sample of elementary school students' ratings of science education before the intervention that the three-dimensional model fit best at both levels of analysis. Table 1 also shows negative relationships at both levels of analysis between students' initial level of science competence and language proficiency and their later obtained scores of cognitive activation and supportive climate. Negative correlations between students' preconditions for learning and classroom process quality are a well-known result in empirical research (e.g., Anderson, Ryan, & Shapiro, 1989; Klieme et al., 2008). At the classroom level, teachers adapt instruction and enhance their engagement in responding particularly to students with low levels of competence (e.g., Klieme et al., 2008). At the individual level, students at risk in particular are supposed to receive more teacher support and thus to report higher levels of classroom process quality (e.g., for cognitive activation, "Our teacher asks me what I have understood and what I haven't"; and for supportive climate, "Our teacher compliments me when I did something good").

Hypotheses 1 and 2: Effects of Embedded Formative Assessment and Classroom Process Quality on Students' Science Understanding

Regarding Hypothesis 1, results showed that students' average level of science understanding in the embedded formative assessment group was higher than in the control group, in which students were taught the same unit on floating and sinking but received no further specific instructions (Table 2, Model 1; see also Decristan et al., 2015). For the present article, we also examined the predictive power of each of the three dimensions of classroom process quality on students' science understanding (Table 2, Models 2 to 4). As expected in Hypothesis 2, cognitive activation, supportive climate, and classroom management each had positive effects on students' learning outcomes.

Hypothesis 3: Interaction Effects Between Embedded Formative Assessment and Classroom Process Quality

Finally and most importantly, we examined the moderating effects of both cognitive activation and supportive climate on treatment effectiveness (Table 3, Models 1 and 2). The results confirmed Hypothesis 3 and revealed positive interactions between classroom process quality and the treatment variable: The effect of embedded formative assessment on students' science understanding, when compared with the control group, increased with higher levels of cognitive activation or supportive climate. The effect sizes at the class level of both interaction models were large. At the same time, and as expected, the results showed no interaction effect between classroom management and the treatment variable on students' science understanding

Table 2
Multilevel Regression Analysis Predicting Students' Science Understanding From the Treatment (Embedded Formative Assessment) and Student Ratings of Classroom Process Quality

	Model 1		Model 2		Model 3		Model 4	
	B	*SE*	*B*	*SE*	*B*	*SE*	*B*	*SE*
Level 1: Control variables								
Cognitive ability	0.15**	0.04	0.15**	0.04	0.15**	0.04	0.14**	0.04
Science competence	0.25**	0.04	0.25**	0.04	0.25**	0.04	0.25**	0.04
Language proficiency	0.13**	0.05	0.13**	0.05	0.13**	0.05	0.13**	0.05
Level 2: Class-level predictors								
Treatment	0.20*	0.11						
Cognitive activation			0.09*	0.05				
Supportive climate					0.11*	0.05		
Classroom management							0.12*	0.07
R^2 (Level 1)	.128		.130		.130		.125	
R^2 (Level 2)	.115		.097		.132		.162	

Note. The treatment variable is dummy coded (control group = 0, embedded formative assessment = 1).
*$p < .05$. **$p < .01$, one-tailed test.

(Table 3, Model 3). Figure 1 shows the role of each dimension of classroom process quality in the effectiveness of the treatment.

Discussion

In this study we examined the interplay between global factors of classroom process quality and curriculum-embedded formative assessment, a well-known teaching practice, in promoting elementary school students' science understanding. To this end, we used data from a cluster-randomized controlled trial with standardized intervention conditions. We employed multilevel regression analysis to examine the main and interaction effects of embedded formative assessment and aggregated student ratings of classroom process quality on students' science understanding. First, as previously presented and discussed by Decristan et al. (2015), embedded formative assessment is an effective tool to enhance elementary students' science understanding (Hypothesis 1). Meta-analyses have revealed the value of formative assessment for student learning (e.g., Hattie, 2009; Kingston & Nash, 2011). However, in this study we demonstrated that the specific contribution of embedded formative assessment is over and above the effect of merely teaching content matter: An inquiry-based science unit was conducted

Table 3

Multilevel Regression Analysis Predicting Students' Science Understanding From the Treatment (Embedded Formative Assessment), Student Ratings of Classroom Process Quality, and Their Interactions

	Model 1		Model 2		Model 3	
	B	*SE*	*B*	*SE*	*B*	*SE*
Level 1: Control variables						
Cognitive ability	0.15**	0.04	0.15**	0.04	0.14**	0.04
Science competence	0.26**	0.04	0.26**	0.04	0.25**	0.04
Language proficiency	0.13**	0.05	0.13**	0.05	0.13**	0.05
Level 2: Class-level predictors						
Treatment	0.19*	0.10	0.18*	0.10	0.17	0.11
Cognitive activation	–0.06	0.03				
Supportive climate			–0.08	0.05		
Classroom management					0.10*	0.06
Interactions between class-level variables						
Treatment × cognitive activation	0.25**	0.08				
Treatment × supportive climate			0.30**	0.08		
Treatment × classroom management					0.02	0.12
R^2 (Level 1)	.135		.135		.123	
R^2 (Level 2)	.437		.629		.206	

Note. The treatment variable is dummy coded (control group = 0, embedded formative assessment = 1).
*$p < .05$. **$p < .01$, one-tailed test.

with both an experimental group and a control group and supported student learning (Decristan et al., 2015; Hardy at al., 2006).

Next, we expected aggregated elementary school student ratings of the three dimensions of classroom process quality (i.e., cognitive activation, supportive climate, and classroom management) to be connected positively with students' science understanding (Hypothesis 2). We showed that, for the larger sample with both intervention conditions, each of the three dimensions of classroom process had a positive effect on students' science understanding, which is consistent with findings from previous research (Klieme et al., 2009; Lipowsky et al., 2009; Pianta et al., 1997; Pianta & Hamre, 2009; Reyes et al., 2012).

Finally, the quality of enactment of embedded formative assessment in class can be linked to global factors of classroom process quality (i.e., cognitive activation and supportive climate; Klieme et al., 2009; Pianta et al., 2008) and thus to principles of effective teaching (see Good et al., 2009). Consistent with research on the interplay between quality components and structural components on learning outcomes (Landenberger & Lipsey, 2005;

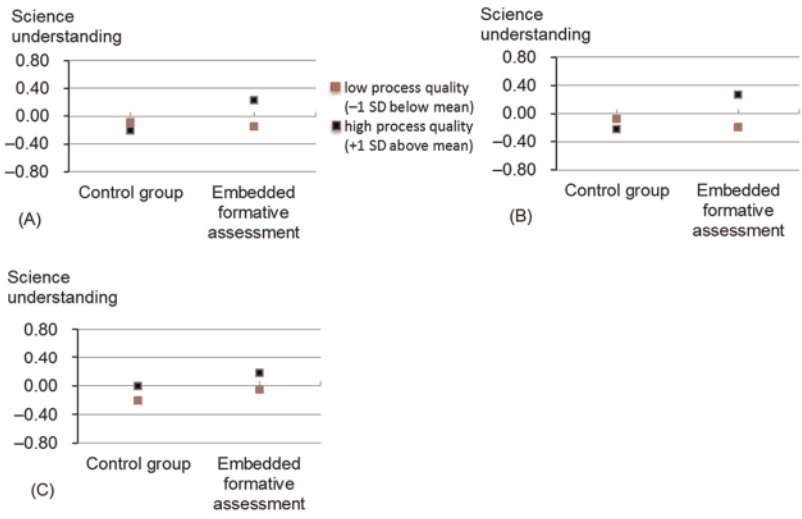

Figure 1. **The moderating effects of classroom process quality on treatment effectiveness. A = cognitive activation; B = supportive climate; C = classroom management.**

O'Donnell, 2007), we expected classroom process quality (i.e., cognitive activation and supportive climate) to enhance the effectiveness of embedded formative assessment (Hypothesis 3). To date, there has not been much work performed on interaction effects involving embedded formative assessment. The results of our study confirmed that high levels of cognitive activation or a supportive climate combined with embedded formative assessment had the most positive effect on students' science understanding. The two corresponding interaction models (Table 3, Models 1 and 2) demonstrated large effects at the classroom level. In contrast, results showed no interaction effect between classroom management and embedded formative assessment on students' science understanding (Table 3, Model 3).

Our findings indicate that the quality enactment of embedded formative assessment (i.e., diagnostic tasks, student feedback, and adapted instruction) is vital for its effectiveness. For instance, diagnostic tests should elicit students' current understanding, and information gathered should correspond with what is taught in the curriculum and should be interpreted correctly by the teachers for further adaptation of instruction (Wilson & Sloane, 2000). Furthermore, formal feedback should provide students with information on their current conceptual understanding, their learning progress and learning goals to help and encourage students to take the next learning step (Hattie & Timperley, 2007; Sadler, 1989).

It should be noted that in our study the diagnostic tests and semistructured student feedback sheets given to the teachers had been designed for implementation in the curriculum. The standardized materials for the use of embedded formative assessment in class had been developed to examine students' current level of conceptual understanding and to guide future teaching and learning. Nevertheless, the effectiveness of the intervention varied with the quality of classroom processes. Thus, it can be speculated that training teachers in the use of embedded formative assessment in class and providing them with high-quality materials is necessary but not sufficient to ensure appropriate use in the classroom. Rather, our results indicate that the effectiveness of embedded formative assessment for student learning depends on how the teacher supports students and how he or she keeps them cognitively active during lessons. Thus, compared with the control group, embedded formative assessment is most effective when it is implemented with high levels of cognitive activation and student support—dimensions of classroom process quality that are largely based on an alignment of teacher prompts and feedback with content-specific student learning trajectories and progressions (e.g., Alonzo & Gotwals, 2012; Duschl, Maeng, & Sezen, 2011, in the context of science education). These global factors of classroom process quality are not necessarily improved by providing assessment tools and teacher training in formative assessment practices. Rather, students' science understanding is enhanced best when the specific teaching practice (embedded formative assessment) is combined with high-quality classroom processes (supportive climate and cognitive activation). In addition, further content-related factors should add to the effectiveness of teaching, as for instance expressed by Good et al. (2009). These factors were not considered in the present study, where content was standardized for all participating classes. However, there is ample evidence that covering content—sometimes referred to as opportunity to learn (OTL; see Schmidt & Maier, 2009)—is a strong predictor of student learning, which is conceptually independent from teaching practices or classroom process quality.

It also should be noted that we tested interactions between individual-level covariates with second-level predictor variables in each of our multilevel analyses to account for interactions between prior group differences and classroom process quality and the treatment variable. Results showed a significant cross-level interaction between student cognitive ability and supportive climate only. Furthermore, considering all cross-level interactions as covariates in our regression models does not change the main conclusions we draw from the study.

The importance of both teacher evaluation and teacher training is obvious. For instance, when designing professional development workshops, trainers should combine specific teaching practices with strategies to enhance the quality of enacting those strategies, and a focus on global factors of classroom process quality. Future studies should explore teacher training (e.g., duration, emphasis on the quality of implementing training

content in class) and teacher characteristics (e.g., beliefs, professional experience) that foster or hinder effective teaching. For instance, an approach to a broader understanding of teacher development that also takes teachers' lifetime experiences into account is to consider teachers' "embodied understanding of practice" (Dall'Alba & Sandberg, 2006). Embodied understanding refers to the way teachers understand their practice (e.g., as knowledge transfer or facilitating learning) which in turn affects teachers' acquisition of knowledge and development of skills. To add further to research on teaching effectiveness, future research should examine the interplay between classroom process quality and other teaching practices as well as the particular connection between global factors of high-quality teaching and quality components of specific teaching practices (e.g., quality of feedback given within formative assessment; see Rakoczy et al., 2013). Finally, research needs to address a variety of content areas to understand the generalizability of factors leading to effective teaching, and their connection to content-related factors such as OTL.

Study Limitations

We examined the specific effects of three dimension of classroom process quality on students' science understanding in 28 classes. Because the sample size at the classroom level was fairly small for multilevel analysis, we were not able to integrate the three parallel models (Table 2, Models 2 to 4), or thus to estimate the relative contribution of each dimension compared with the other two dimensions for student learning.

The present article focused on the interplay between class-level variables (components of effective teaching) while controlling for students' proximal variables of science understanding (i.e., cognitive ability, science competence, and language proficiency). However, we did not consider more distal variables (e.g., immigrant background, socioeconomic status) that could affect students' science understanding. To provide insight into second language learners' science understanding, the role of culture and socioeconomic status, as well as students' proficiency in their first and second languages should be investigated (for a review on this topic, see Lee, 2005).

In this article, global factors of effective teaching were assessed based on students' ratings of classroom process quality. Although the value of student ratings has been challenged (e.g., Greenwald, 1997), Benton and Cashin (2012) point out that there has been "more than 50 years of credible research on the validity and reliability of student ratings" (p. 2). Recent research has shown that even elementary school students' ratings can be used as a reliable and valid measure of classroom process quality (Allen & Fraser, 2007; Doll et al., 2010; Fauth et al., 2014). However, for certain features of classroom instruction, such as cognitive activation, additional examination of the perspectives of observers and teachers may provide further insight into the links between

teaching practices, perspectives on classroom process quality, and student learning outcomes (e.g., Benton & Cashin, 2012; Kunter & Baumert, 2006).

Finally, we did not administer or examine the results of a test of science understanding before conducting our analysis. Although several studies have provided evidence for the validity of the items on the test we administered (Hardy et al., 2006; Pollmeier et al., in press; Schneider & Hardy, 2013), particularly preinstructional concepts often do not follow the order of item difficulty assumed for experts or persons more involved conceptually in the topic. Rather the conceptual understanding of floating and sinking systematically develops through exposure to the content during the curriculum, and students then develop a more consistent view, resulting in more reliable test scores at the end of the curriculum (see Decristan et al., 2015).

Conclusions

The results of our study show that students' science understanding improves through embedded formative assessment when combined with sufficient levels of global factors of classroom process quality (supportive climate and cognitive activation). This empirical finding may help scholars find a compromise between the two general strands of research on teaching effectiveness outlined at the beginning of our article, namely specific teaching practices versus global aspects of classroom processes. Effective teaching cannot be measured by either checking for a limited set of "best practices" or rating global aspects of classroom process quality only. Contrary to popular belief, no single, reductionist approach to effective teaching will be sufficient. Instead, both can enhance student learning to some extent, but best results may depend on combining specific and global principles of teaching.

Notes

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

[1]The total intervention included 54 teachers and 1,070 students in three instructional conditions and a control group. The present article extends previous research on treatment effectiveness. It focuses on embedded formative assessment, as it was the only instructional condition that enhanced students' science understanding when compared with the control group (see Decristan et al., 2015).

[2]For organizational reasons, we could not assess the treatment scores in one control group class.

References

Allen, D., & Fraser, B. J. (2007). Parent and student perceptions of classroom learning environment and its association with student outcomes. *Learning Environments Research, 10*, 67–82. doi:10.1007/s10984-007-9018-z

Alonzo, A., & Gotwals, A. (2012). *Learning progressions in science: Current challenges and future directions*. Rotterdam, Netherlands: Sense.

Anderson, L. W., Ryan, D. W., & Shapiro, B. J. (1989). *The IEA Classroom Environment Study*. Oxford, UK: Pergamon.

Ball, D. L., & Forzani, F. M. (2011). Building a common core for learning to teach, and connecting professional learning to practice. *American Educator, 35*(2), 17–21.

Baumert, J., Kunter, M., Blum, W., Brunner, M., Voss, T., Jordan, A., . . . Tsai, Y.-M. (2010). Teachers' mathematical knowledge, cognitive activation in the classroom, and student progress. *American Educational Research Journal, 47*, 133–180. doi:10.3102/0002831209345157

Bell, B., & Cowie, B. (2001). The characteristics of formative assessment in science education. *Science Education, 85*, 536–553. doi:10.1002/sce.1022

Benton, S. L., & Cashin, W. E. (2012). *Student ratings of teaching: A summary of research and literature* (IDEA Paper No. 50). Manhattan, KS: IDEA Center.

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

Bolger, N., Stadler, G., & Laurenceau, J. P. (2012). Power analysis for intensive longitudinal studies. In M. R. Mehl & T. S. Conner (Eds.), *Handbook of research methods for studying daily life* (pp. 285–301). New York, NY: Guilford.

Bransford, J. D., Brown, A. L., & Cocking, R. R. (2000). *How people learn*. Washington, DC: National Academy Press.

Brophy, J. (2000). *Teaching*. Brussels, Belgium: International Academy of Education.

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*, 330–351. doi:10.1037//1082-989X.6.4.330-351

Cornelius-White, J. (2007). Learner-centered teacher-student relationships are effective: A meta-analysis. *Review of Educational Research, 77*, 113–143. doi:10.3102/003465430298563

Curby, T. W., Rimm-Kaufman, S. E., & Ponitz, C. C. (2009). Teacher-child interactions and children's achievement trajectories across kindergarten and first grade. *Journal of Educational Psychology, 101*, 912–925. doi:10.1037/a0016647

Dall'Alba, G., & Sandberg, J. (2006). Unveiling professional development: A critical review of stage models. *Review of Educational Research, 76*, 383–412. doi:10.3102/00346543076003383

Danielson, C. (1996). *Enhancing professional practice: A framework for teaching*. Alexandria, VA: Association of Supervision and Curriculum Development.

Davis, H. A. (2003). Conceptualizing the role and influence of student-teacher relationships on children's social and cognitive development. *Educational Psychologist, 38*, 207–234. doi:10.1207/S15326985EP3804_2

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. *Journal of Educational Research*. Advance online publication. doi:10.1080/00220671.2014.899957

Doll, B., Spies, R. A., LeClair, C. M., Kurien, S. A., & Foley, B. P. (2010). Student perceptions of classroom learning environments: Development of the ClassMaps survey. *School Psychology Review, 39*, 203–218.

Doyle, W. (1986). Classroom organization and management. In M. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 392–431). New York, NY: Macmillan.

Duschl, R., Maeng, S., & Sezen, A. (2011). Learning progressions and teaching sequences: A review and analysis. *Studies in Science Education*, *47*, 123–182. doi:10.1080/03057267.2011.604476

Elben, C. E., & Lohaus, A. (2001). *Marburger Sprachverständnistest für Kinder ab 5 Jahren* [Marburger test of language comprehension for children 5 years of age and older]. Göttingen, Germany: Hogrefe.

Emmer, E. T., & Stough, L. M. (2001). Classroom management: A critical part of educational psychology, with implications for teacher education. *Educational Psychologist*, *36*, 103–112. doi:10.1207/S15326985EP3602_5

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods*, *12*, 121–138. doi:10.1037/1082-989X.12.2.121

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., 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

Glück, C. W. (2011). *Wortschatz- und Wortfindungstest für 6– bis 10–Jährige: WWT 6–10* [Test of vocabulary of 6- to 10-year-old children]. Munich, Germany: Elsevier, Urban & Fischer.

Good, T. L., Wiley, C. R. H., & Florez, I. R. (2009). Effective teaching: An emerging synthesis. In L. J. Saha & A. G. Dworkin (Eds.), *International handbook of research on teachers and teaching* (pp. 803–816). New York, NY: Springer.

Greenwald, A. G. (1997). Validity concerns and usefulness of student ratings of instruction. *American Psychologist*, *52*, 1182–1186. doi:10.1037/0003-066X.52.11.1182

Grossman, P., Loeb, S., Cohen, J., & Wyckoff, J. (2013). Measure for measure: The relationship between measures of instructional practice in middle school English language arts and teachers' value-added. *American Journal of Education*, *119*, 445–470. doi:10.1086/669901

Gustafsson, J.-E., & Balke, G. (1993). General and specific abilities as predictors of school achievement. *Multivariate Behavioral Research*, *28*, 407–434. doi:10.1207/s15327906mbr2804_2

Gwet, K. L. (2012). *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among multiple raters*. Gaithersburg, MD: Advanced Analytics Press.

Hamre, B. K., & Pianta, R. C. (2005). Can instructional and emotional support in the first-grade classroom make a difference for children at risk of school failure? *Child Development*, *76*, 949–967. doi:10.1111/j.1467-8624.2005.00889.x

Hamre, B. K., Pianta, R. C., Mashburn, A. J., & Downer, J. T. (2007). *Building a science of classrooms: Application of the CLASS framework in over 4,000 U.S. early childhood and elementary classrooms*. New York, NY: Foundation for Child Development.

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

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

Hertel, S. (2009). *Beratungskompetenz von Lehrern—Kompetenzdiagnostik, Kompetenzförderung, Kompetenzmodellierung* [Teachers' parental counseling competency—Diagnosis, support, and modelling of competency]. Münster, Germany: Waxmann.

Hondrich, A. L., Hertel, S., Adl-Amini, K., & Klieme, E. (in press). Implementing curriculum-embedded formative assessment in primary school science classrooms. *Assessment in Education: Principles, Policy & Practice*.

Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013). *Have we identified effective teachers? Validating measures of effective teaching using random assignment* (Research paper, MET Project). Seattle, WA: Bill & Melinda Gates Foundation.

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., Jude, N., Rauch, D., Ehlers, H., Helmke, A., Eichler, W., . . . Willenberg, H. (2008). Alltagspraxis, Qualität und Wirksamkeit des Deutschunterrichts [Everyday practice, quality, and effectiveness of German lessons.] In DESI-Konsortium (Eds.), *Unterricht und Kompetenzerwerb in Deutsch und Englisch. Ergebnisse der DESI-Studie* (pp. 319–344). Weinheim, Germany: Beltz.

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, Germany: Waxmann.

Kounin, J. S. (1970). *Discipline and group management in classrooms*. New York, NY: Holt, Rinehart & Winston.

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

Kunter, M., Klusmann, U., Baumert, J., Richter, D., Voss, T., & Hachfeld, A. (2013). Professional competence of teachers: Effects on quality and student development. *Journal of Educational Psychology*, *105*, 805–820. doi:10.1037/a0032583

Landenberger, N. A., & Lipsey, M. W. (2005). The positive effects of cognitive-behavioral programs for offenders: A meta-analysis of factors associated with effective treatment. *Journal of Experimental Criminology*, *1*, 451–476. doi:10.1007/s11292-005-3541-7

Lee, O. (2005). Science education with English language learners: Synthesis and research agenda. *Review of Educational Research*, *75*, 491–530. doi:10.3102/00346543075004491

Lipowsky, F., Rakoczy, K., Pauli, C., Drollinger-Vetter, B., Klieme, E., & Reusser, K. (2009). Quality of geometry instruction and its short-term impact on students' understanding of the Pythagorean theorem. *Learning and Instruction*, *19*, 527–537. doi:10.1016/j.learninstruc.2008.11.001

Lüdtke, O., Robitzsch, A., Trautwein, U., & Kunter, M. (2009). Assessing the impact of learning environments: How to use student ratings of classroom or school characteristics in multilevel modeling. *Contemporary Educational Psychology*, *34*, 120–131. doi:10.1016/j.cedpsych.2008.12.001

Lüdtke, O., Trautwein, U., Kunter, M., & Baumert, J. (2006). Reliability and agreement of student ratings of the classroom environment: A reanalysis of TIMSS data. *Learning Environments Research*, *9*, 215–230. doi:10.1007/s10984-006-9014-8

Martin, M. O., Mullis, I. V. S., & Foy, P. (with Olson, J. F., Erberber, E., Preuschoff, C., & Galia, J.). (2008). *TIMSS 2007 international science report: Findings from*

IEA's Trends in International Mathematics and Science Study at the fourth and eighth grades. Chestnut Hill, MA: Boston College, TIMSS & PIRLS International Study Center.

Martin, M. O., Mullis, I. V., Foy, P., & Stanco, G. M. (2012). *TIMSS 2011 international results in science*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149–174. doi:10.1007/BF02296272

Möller, K., & Jonen, A. (2005). *Die KiNT-Boxen-Kinder lernen Naturwissenschaft und Technik. Paket 1: Schwimmen und Sinken* [The KiNT-boxes—Children learn science and technology. Package 1: Floating and sinking]. Essen, Germany: Spectra.

Muthén, L. K., & Muthén, B. O. (1998–2013). *Mplus* (Version 7.11) [Computer software]. Los Angeles, CA: Muthén & Muthén.

O'Donnell, C. L. (2007). *Fidelity of implementation to instructional strategies as a moderator of curriculum unit effectiveness in a large-scale middle school science quasi-experiment* (Doctoral dissertation). Retrieved from Dissertation Abstracts International. (UMI No. AAT 3276564)

Osborne, J. F., Erduran, S., & Simon, S. (2004). Enhancing the quality of argument in school science. *Journal of Research in Science Teaching*, *41*, 994–1020. doi:10.1002/tea.20035

Petermann, F., Metz, D., & Fröhlich, L. P. (2010). *SET 5–10. Sprachstandserhebung stest für Kinder im Alter zwischen 5 und 10 Jahren* [Test of language comprehension of 5- to 10-year-old children]. Göttingen, Germany: Hogrefe.

Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher*, *38*, 109–119. doi:10.3102/0013189X09332374

Pianta, R. C., La Paro, K., & Hamre, B. K. (2008). *Classroom Assessment Scoring System (CLASS)*. Baltimore, MD: Brookes.

Pianta, R. C., Nimetz, S. L., & Bennett, E. (1997). Mother-child relationships, teacher-child relationships, and school outcomes in preschool and kindergarten. *Early Childhood Research Quarterly*, *12*, 263–280. doi:10.1016/S0885-2006(97)90003-X

Pollmeier, J., Troebst, S., Hardy, I., Möller, K., Kleickmann, T., Jurecka, A., & Schwippert, K. (in press). Science-P I: Modeling conceptual understanding in primary school. In D. Leutner, J. Fleischer, J. Grünkorn, & E. Klieme (Eds.), *Competence assessment in education: Research, models and instruments*. New York, NY: Springer.

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. doi:10.1016/j.learninstruc.2013.03.002

Raudenbush, S. W. (2008). Advancing educational policy by advancing research on instruction. *American Educational Research Journal*, *45*, 206–230. doi:10.3102/0002831207312905

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks, CA: Sage.

Reyes, M. R., Brackett, M. A., Rivers, S. E., White, M., & Salovey, P. (2012). Classroom emotional climate, student engagement, and academic achievement. *Journal of Educational Psychology*, *104*, 700–712. doi:10.1037/a0027268

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, *18*, 119–144. doi:10.1007/BF00117714

Scheerens, J., & Bosker, R. J. (1997). *The foundations of educational effectiveness*. Oxford, UK: Pergamon.

Schmidt, W. H., & Maier, A. (2009). Opportunity to learn. In G. Sykes, B. L. Schneider, & D. N. Plank (Eds.), *Handbook on education policy research* (pp. 541–549). New York, NY: Routledge.

Schneider, M., & Hardy, I. (2013). Profiles of inconsistent knowledge in children's pathways of conceptual change. *Developmental Psychology*, *49*, 1639–1649. doi:10.1037/a0030976

Schütte, K., Frenzel, A. C., Asseburg, R., & Pekrun, R. (2007). Schülermerkmale, naturwissenschaftliche Kompetenz und Berufserwartung. [Student characteristics, science competence, and career expectation.] In PISA-Konsortium Deutschland (Ed.), *PISA 2006–Die Ergebnisse der dritten internationalen Vergleichsstudie* (pp. 125–146). Münster, Germany: Waxmann.

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

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

Shemwell, J., & Furtak, E. (2010). Science classroom discussion as scientific argumentation: A study of conceptually rich (and poor) student talk. *Educational Assessment*, *15*, 222–250. doi:10.1080/10627197.2010.530563

Wang, M. C., Haertel, G. D., & Walberg, H. J. (1993). Toward a knowledge base for school learning. *Review of Educational Research*, *63*, 249–294. doi:10.3102/00346543063003249

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*, 427–450. doi:10.1007/BF02294627

Weiß, R. H. (2006). *CFT 20–R. Grundintelligenztest Skala 2—Revision* [Culture Fair Test]. Göttingen, Germany: Hogrefe.

Wilson, M., & Sloane, K. (2000). From principles to practice: An embedded assessment system. *Applied Measurement in Education*, *13*, 181–208. doi:10.1207/S15324818AME1302_4

Manuscript received April 28, 2014
Final revision received October 8, 2014
Accepted March 10, 2015