Zitiervorschlag: Smit, R., Bachmann, P., Blum, V., Birri, T., & Hess, K. (2017). Effects of a rubric for mathematical reasoning on teaching and learning in primary school. *Instructional Science*, 45(5), 603-622.
<https://doi.org/10.1007/s11251-017-9416-2>

Zur Verfügung gestellt auf PHIQ:

PHIQ-DOI: <https://doi.org/10.18747/PHSG-coll3/id/505>
Original-DOI: <https://doi.org/10.1007/s11251-017-9416-2>

Dokumentart: Journal Article

Version: accepted version

Copyright-Hinweis: This is a post-peer-review, pre-copyedit version of an article published in *Instructional Science*. The final authenticated version is available online at: <https://doi.org/10.1007/s11251-017-9416-2> .

Lizenz: Alle Rechte vorbehalten

Effects of a rubric for mathematical reasoning on teaching and learning in primary school

Robbert Smit¹; Patricia Bachmann¹; Verena Blum²; Thomas Birri¹; Kurt Hess²

Abstract Rubrics are assessment tools that help students gain complex competencies. Our quasi-experimental study aimed to evaluate whether rubrics help teachers teach and assess mathematical reasoning in primary school and whether such an instrument might support student learning. In two Swiss cantons, 762 students in 44 5th- and 6th-grade primary classes worked on their reasoning competencies, and half of them additionally employed our standards-based rubric. All of the teachers received a 1-day training and participated in the final project evaluation. To standardise and support the teachers during the implementation phase, they received a detailed curriculum. An achievement test and questionnaires for students and teachers were administered before and at the end of the intervention. The results of our quantitative longitudinal analyses indicate that the rubric fosters the teachers' perceived diagnostic skills but only indirectly impacts their use of formative feedback. Based on the students' perceptions, however, we observed a direct effect of the rubric on formative feedback and student self-assessment. Effects on students' outcomes could not be observed, but there are indications of effects mediated by self-regulation and self-efficacy.

Keywords Rubrics; Formative assessment; Mathematical reasoning; Primary school; Feedback

¹ University of Teacher Education St. Gallen, St. Gallen, Switzerland

² University of Teacher Education Zug, Zug, Switzerland

Introduction

A rubric is a coherent set of criteria for students' work that includes descriptions of levels of performance quality (Brookhart 2013). It can be used for summative or formative assessment purposes. The students' learning is supported by a rubric through the clarifying of learning targets, giving of teacher- or peer-guided feedback and student self-assessment. Teachers can benefit from rubrics as a tool to specifically diagnose the learning level and process of the students in the classroom (Ash and Levitt 2003). A rubric not only can guide a teacher's feedback, but it can also serve as a supporting instrument for peer- and self-assessment as part of the learning process (Panadero et al. 2016; Saddler and Andrade 2004). Linked to this, the students might develop their self-regulation competencies and their attitudes of self-efficacy (Panadero and Jonsson 2013).

Opportunities for self-regulated learning are embedded in tasks that are complex by design (Perry et al. 2004), e.g., mathematical problem solving or reasoning and proof. Using rubrics in classroom assessment for complex skills, such as essay writing or in our case mathematical reasoning, might ensure that teaching covers all the standards and not just the easily measurable ones. In the Framework for Classroom Assessment in Mathematics (de Lange 1999), it is expected that teachers support students in solving problems, posing questions, and communicating using mathematical language. As part of this framework, teachers should design and use higher-level tasks combined with instructionally embedded assessment. This practice includes providing learning opportunities for the students to identify their own academic strengths, determine their areas of need, and monitor their progress in achievement (Bransford and Donovan 2005). Over the past decade, there has been increased recognition of the importance of mathematical reasoning in students' mathematical education. For example, in the United States, the National Council of Teachers of Mathematics (NCTM) (2000) and, in Germany the Standing Conference of the Ministers of Education and Cultural Affairs of the Länder in the Federal Republic of Germany (KMK) 2004, have made calls for reasoning and proving to become central to all students' mathematical experiences across all school years. Although in most elementary mathematics problems, the student's task is to discover the pattern of relations or attributes among several elements given in a problem, the stimulation of thinking skills is not pursued explicitly (Christou and Papageorgiou 2007). It is usually assumed that these skills develop as a by-product of the teaching of content as defined in traditional curricula for different subjects (Hamers et al. 1998). Not surprisingly, many students have difficulties in solving mathematical problems, as international studies such as PISA 2003 (Organization for Economic Co-Operation and Development (OECD) 2004) and TIMSS 2003 (Mullis et al. 2004) have demonstrated. First results from PISA 2012 (Organization for Economic Co-Operation and Development (OECD) 2014) indicate that only 20% of students in OECD countries frequently encounter mathematics problems that are set in real-world context and where argumentation skills are demanded.

A significant amount of literature is dedicated to the practical use of rubrics in primary and secondary schools, and many teachers produce rubrics themselves. Experimental research on the effects of rubrics for learning is still scarce (Brookhart and Chen 2015), particularly for science and mathematics (Andrade et al. 2010). Our research adds to the understanding of how rubrics support student learning by following a research model with mediating variables oriented according to a theoretical model developed by Panadero and Jonsson (2013). The inclusion of a multitude of variables allows for a more valid modelling of complex phenomena (Sandoval and Bell 2004).

The current study first outlines the function and the effects of rubrics. Next, it provides a short introduction into mathematical reasoning, the construction of our rubric, and its relation to formative assessment practices, such as giving feedback and peer- and self-assessment.

Theoretical background

What are rubrics and how are they implemented in the classroom?

Rubrics help teachers evaluate complex competencies assessed in authentic performance tests, such as written compositions, oral presentations, or science projects (Libbee 2001; Popham 1997). The instrument articulates the expectations for an assignment by listing criteria and describing levels of quality (Andrade 2000). Rubrics can be holistic or analytic; holistic rubrics provide a single score based on an overall impression of a student's performance on a task, whereas analytic rubrics (see our rubric in Smit and Birri 2014) provide specific feedback on several dimensions and levels. To use a rubric as a teaching tool, strategies for its application in the classroom are needed (see e.g. Arter and Chappuis 2007). Hsia et al. (2015, p. 5) arrange five typical strategies in a circle model with different steps. In step 0, as part of the lesson planning, the teacher might use the rubric to choose tasks related to the learning goals and the dimensions of the rubric. In the classroom, step 1 would be to clarify the learning targets to the students. In the second step, the teacher explains the rubric in detail and provides some examples to illustrate the different levels of the criteria. The third step consists in the students gaining practice applying the rubric to check their work. At the same time, the teacher obtains detailed diagnostic information as a basis for day-to-day instructional decisions. Based on these diagnoses, the teacher gives feedback according to the dimensions and levels of the instrument (step four). Finally (step five), the students reach a level at which they demonstrate their final performance. Accordingly, the teacher either asks the students to revise their work based on the rubric or assesses the students' cumulative performance and begins a new topic.

Effects of rubrics on teaching and on students' learning

Rubrics provide a good framework for a teacher to give feedback on a student's actual performance level as well as to indicate the next steps for improvement. Handing out the rubric to make goals transparent is not enough or not as effective as providing information for individual improvement (Wollenschläger et al. 2016). In addition, teachers with limited perceptions of teaching and learning, e.g., strong transmissive beliefs of learning, might hinder the full potential of a rubric for students' learning (Mui So and Hoi Lee 2011). Effective formative feedback depends on teachers' diagnostic skills (Earl 2012; Turner 2014). Ideally, teachers apply their diagnostic skills not only when devising, correcting, and grading tests and examinations but particularly when preparing lessons and monitoring students' understanding during the learning process (Brunner et al. 2013); both teacher activities are part of formative assessment. Providing feedback in classrooms has been found to have a powerful impact on students' learning, with an overall effect size of $d = 0.79$ (Hattie 2008). In general and not exclusive to rubrics, feedback that encourages 'mindfulness' is most likely to help students improve (Hattie and Timperley 2007; Underwood and Tregidgo 2006). That is, comments that prompt students to meaningfully

and thoughtfully approach revisions tend to result in the highest gains in performance. Bangert-Drowns et al. (1991) found that although feedback was positively related to greater achievement in most settings, student performance did not improve if feedback messages failed to include information necessary for learners to evaluate where they are and where they are going or did not provide useful strategies to get them there. Rubrics can provide information for all three conditions and help the teacher to keep feedback simple. The complexity of feedback information also seems to reduce the effectiveness of formative feedback (Shute 2008).

Rubrics are not only a supportive tool for the teacher but also a helpful instrument for the student. In their review on the impact of rubrics, Panadero and Jonsson (2013) found that the assessment tool helps students to improve their self-efficacy, reduces anxiety, and fosters their ability to self-regulate their learning. Feedback is generally seen as an inherent catalyst for all self-regulated activities (Butler and Winne 1995). Andrade et al. (2008) showed that students in elementary classes using a rubric to self-assess first drafts of an essay could amend the quality of their writing. This research group positively replicated their results with middle school students (Andrade et al. 2010). In a study by Panadero et al. (2012), students at the end of secondary school using a rubric achieved higher self-regulation competencies than those in the control group. The rubric group also developed higher self-efficacy beliefs than the control group in combination with process related feedback provided during the intervention.

With respect to the review of the positive effects of formative rubrics by Panadero and Jonsson (2013), it becomes evident that the effects of rubrics on student outcomes or self-regulation are often not direct but are mediated in different ways, e.g., by student self-efficacy or through meta-cognitive activities (Panadero and Jonsson 2013; Phye and Sanders 1994). Perry and Winne (2006) developed a research-based model of self-regulated learning. The model demonstrates how students' self-regulation depends on external feedback from peers and the teacher. A rubric supports the students' metacognitive monitoring such that they can compare the features of a current product—the target of monitoring—to a list of criteria that describe the qualities or properties of the goal. In addition to monitoring, self-efficacy is also related to self-regulation (Zimmerman 2008). Self-efficacy influences goals a student sets, commitment to those goals, decision making at branch points along a path the learner constructs to reach those goals, and persistence (Bandura 1993). Self-efficacy but not self-regulation was a predictor for mathematic achievement in a study on motivational effects with 6th graders by Pajares and Graham (1999), but a strong relation between self-regulation and self-efficacy appeared in their results. Andrade et al. (2009) showed in their research that a rubric fosters students' self-efficacy.

Mathematical reasoning

Currently, mathematical reasoning is viewed as necessary to ensure that students understand mathematics concepts and skills (Thompson and Schultz-Ferrel 2008). Algebraic reasoning in primary school can take various forms, such as exploring patterns and describing relationships. The introduction of algebraic reasoning into primary classrooms requires new competencies of class teachers because most of these teachers have little experience with the rich and connected aspects of algebraic reasoning (Blanton and Kaput 2005). The Swiss mathematical standards for students at the end of primary school (grade six) require, in short, the ability to verify statements and to justify or falsify results using data or arguments. The standard does not include formal proofs, which are not a topic in

primary and secondary school curricula but are addressed in high school. Building argumentation skills as part of pre-formal proof (Blum and Kirsch 1991), however, begins in primary school (Semadeni 1984) and includes explaining procedures, assumptions and results, and making claims, predictions and generalisations (Bezold 2009). Tasks that foster reasoning competence invite the use of multiple-solution strategies and multiple representations and require that students explain or justify how they arrived at their answers (Stein et al. 1996). Teachers must be aware that it is not sufficient to present such tasks without creating an instructional environment with an emphasis on discourse. This practice is important for the teacher's understanding of students' thinking process (Brodie 2010; Ginsburg 2009; Sfard 2001). Because reasoning in mathematical lessons requires interaction with others, a (socio)-constructivist view of learning mathematics could be viewed as more favourable than a transmissive view (Voss et al. 2013). Thus, teachers' beliefs about mathematics, teaching, and learning can constrain the ways tasks are implemented (Swan 2007), and as a consequence, unfavourable beliefs need to be altered. Furthermore, time constraints hinder teachers from introducing contextualised investigations and activities that encourage groups of students to engage in learning mathematical concepts and applications and in reflective writing and talking about these ideas (Keiser and Lambdin 1996). Such activities relate to, e.g., increased group work, writing, extended projects, and alternative forms of assessment. Therefore, appropriate forms of professional support are necessary to change instructional and curricular practices. Children with a foreign language background are known to have more difficulties with word problems than those who are native to a given country (Kempert et al. 2011; Stanat and Christensen 2006). This difficulty may arise because of linguistic structures that imply mathematical structures and operations, for example, understanding relational terms such as 'more than' or 'less than'. Similar problems exist for children with special needs (Kroesbergen and Van Luit 2003). In addition, such children often have inadequate strategies for problem solving (Swanson 2014). Kroesbergen and Van Luit (2003) consider self-instruction methods as effective as direct instruction when it comes to problem solving. To develop strong self-regulation habits in self-instruction learning situations, learners can benefit from metacognitive scaffolds such as teacher prompts (formative feedback) and rubrics (Turner 2014).

As part of our pilot study, we constructed a rubric aligned with the Swiss standards (Smit and Birri 2014). The rubric should allow teachers to document and report growth of complex competencies over an extended period of time, e.g., two school years. Our instrument consists of four dimensions of sound reasoning for primary school (appropriate and comprehensible procedure, correct computations, comprehensible and detailed description and argumentation, illustrations and examples) and four stages (levels) of development.

Research questions

Our primary aim was to demonstrate the effects of a rubric for mathematical reasoning in grades 5 and 6 on teaching and learning. A secondary target was to explore how the use of our rubric relates to the implementation of formative assessment practices by the teachers (Brown et al. 2012; Hattie and Timperley 2007). As a basis of our research, we partly followed the theoretical model of Panadero and Jonsson (2013). In summary, their comprehensive model outlines how the use of a rubric leads to more transparency of the assessment criteria and as a consequence results in a number of beneficial factors related to student achievement, such as more self-efficacy or better self-regulation. The model allows

for studying innovative learning environments in complex classroom settings in the sense of design based research (Sandoval and Bell 2004). We added to Panadero and Jonssons’ model variables on the teacher’s side that relate to the second target mentioned above: the quality of the teachers’ diagnostic skills and the teachers’ formative feedback. In addition, we included several control variables: gender, teacher experience, transmissive beliefs, and student background, among others.

The model depicted in Fig. 1 allows for a number of connected research questions to be raised. In the presented study, we examined two questions that tackle the direct effects of our intervention on aspects of formative assessment. We were ambivalent about effects of the intervention on self-assessment. Both groups of teachers were ordered to conduct a number of student peer- and self-assessments to prevent confusing our intervention with peer- and self-assessment (Brookhart and Chen 2015). However, students might perceive the rubric differently, as in their eyes, self-assessments may be more visible with the rubric than without.

In addition, we tested whether the rubric has an indirect effect on two student competencies: self-regulation and self-efficacy. Finally, we explored the effects of our intervention on students’ achievement in mathematical reasoning. Researchers are uncertain whether rubrics might have a direct effect on student achievement (Brookhart and Chen 2015). Thus, the following hypotheses are formulated:

Hypothesis 1a: There is a positive effect of our intervention on the teachers’ self-reported diagnostic skills and formative feedback practices.

Hypothesis 1b: There are no effects on the teachers’ reported number of self- and peer-assessments, however.

Hypothesis 2: There is a positive effect of our intervention on the students’ perception of self- and peer-assessment and formative feedback.

Hypothesis 3: There is a positive indirect effect of our intervention on students’ self-regulation and self-efficacy.

Hypothesis 4: There is a positive effect of our intervention on students’ outcomes.

Fig. 1 Research model for the effects of a rubric for mathematical reasoning on teaching and learning

Method

Design

The present study was part of the project ‘Learning with rubrics’ and was conducted by two teacher education universities in Switzerland (St. Gallen and Zug). The study began in 2015 and lasted for two years. Its design was quasi-experimental and longitudinal (Table 1). We measured the students’ competence for mathematical reasoning at the beginning of the project phase (T1) before we conducted a 1-day workshop with theoretical input on mathematical reasoning. In addition, the teachers and the students completed a questionnaire on attitudes and related teaching aspects such as feedback quality. The workshop was held at each teacher university separately but by the same team members. Whereas the intervention group received information on the use of rubrics, the control group discussed non-interfering mathematical related content. The intervention followed the workshop, consisting of 9 weeks with one lesson given each week on practicing mathematical reasoning in the classroom. The teachers had to follow a strict script in the form of detailed lesson plans. Our lesson plans had a socio-constructivist orientation, and collaborative group or peer work (e.g., placemat technique) was part of every lesson. In the intervention group, the implementation of a rubric for self- and teacher-assessment was added to the lesson plans. At the end of the intervention, the teachers again participated in a workshop. Then, we measured students’ competencies, attitudes, and the teachers’ perceptions a second time (T2). One part of the second workshop was intended to evaluate the intervention phase, whereas another part was reserved for the input each group missed in the first workshop. Thus, the intervention group was informed on mathematical representation for primary students, and the control group learned about the use of rubrics as part of formative assessment.

Table 1 Research design

T1 test and questionnaire August 2015	Intervention: training and 9 lessons practical implementation	T2 test and questionnaire November 2015	Evaluation November 2015
Teacher questionnaire	IG and CG: teachers participate in workshop for mathematical reasoning. IG receives supplemental input on the use of rubrics	Teacher questionnaire	IG and CG: evaluation of the implementation of mathematical reasoning in the classroom IG receives supplemental input on mathematical representation
Student questionnaire and test for mathematical reasoning	CG receives supplemental input on mathematical representation.	Student questionnaire and test for mathematical reasoning	CG receives supplemental input on the use of rubrics

IG intervention group, *CG* control group

Sample

We recruited our teachers with the help of an advertisement in the local teacher journal, in addition to personal requests. Our sample consisted of 45 full-time teachers from 2 Swiss cantons. Twenty-two were part of the intervention group, and 23 participated in the control group. Twenty-five teachers were women and 20 men; the mean age was 41, and the mean service age was 11 years. Nine teachers managed a 5th grade class, and 18 teachers were responsible for a 6th grade class. The remaining 18 teachers taught multi-grade classes. Four of them even involved their 4th grade students in the project. The allocation to the two groups—intervention and control groups—was partly random, considering an equal distribution of teacher gender, age, grade, and canton (district). During the project period, one teacher dropped out because of a heavy workload involving daily classroom work, and one teacher determined to be a co-teacher of another participating teacher. Thus, we obtained datasets for 44 classes with 23 4th grade, 337 5th grade and 402 6th grade students (totalling 762 students). Approximately 52% of the students were boys and 48% girls. For 25% of all students, German was not the main language spoken at home.

Measurement

Questionnaire

We employed questionnaires, one for the teacher and one for the students. In each of the two measurements, a bundle of items was repeatedly used, complemented by items appropriate only for a single time point. The items were chosen based on the research model shown in Fig. 1. The teacher scales for the presented study included ‘Diagnostic skills’ with 3 items (t1/t2: $a = 0.75/0.71$); ‘Formative feedback’ with 8 items covering task-, process-, self-regulation, and self-level (t1/t2: $a = 0.73/0.67$); and ‘Peer- and self-assessment’ with 4 items (t1/t2: $a = 0.79/0.76$). In addition, ‘Transmissive beliefs’ were measured with 3 items (t1/t2: $a = 0.63/0.71$) as a control variable for mathematical reasoning. The student scales comprised ‘Formative feedback’ with 5 items (t1/t2: $a = 0.68/0.75$), ‘Peer- and self-assessment’ with three items (t1/t2: $a = 0.52/0.62$), ‘Self-regulation competence’ with 7 items (t1/t2: $a = 0.67/0.70$), and ‘Self-efficacy for reasoning’ with 3 items (t1/t2: $a = 0.66/0.68$).

Background variables also included information on the students’ nationality and special needs. The scales for diagnostic skills, formative feedback and peer- and self-assessment were constructed based on existing items (Brown et al. 2012; Smit 2009) or newly developed based on the literature (Hargreaves et al. 2000; Hattie and Timperley 2007). For the transmissive beliefs, items from Rakoczy et al. (2005) were employed. The items for self-regulation competence were partly adapted from Purdie et al. (1996) and partly newly constructed. Finally, the items for self-efficacy were all newly constructed based on the definition by Zimmerman (2000). For most items, a 6-point Likert scale was employed (6 = *absolutely agree*, 5 = *agree*, 4 = *somewhat agree*, 3 = *somewhat disagree*, 2 = *disagree*, and 1 = *absolutely disagree*, or ‘How often does the following occur during math lessons:’ 6 = *always*, 5 = *almost always*, 4 = *often*, 3 = *sometimes*, 2 = *seldom*, and 1 = *never*). Non-fitting items of scales with low Cronbach’s alpha values were eliminated, or in the case of formative feedback, the student scale was restructured. All items are presented as Supplemental material.

Mathematical reasoning test

All items for measuring mathematical reasoning were either adapted from other standards-based tests or newly developed (see examples in Supplementary material). The items were aligned with the Swiss national basic competences (Swiss Conference of Cantonal Ministers of Education (EDK) 2011). One of the authors was a member of the standard development team. The complete test battery consisted of 18 items, which we distributed over 2 testlets of 10 items each. Thus, most items were used repeatedly to function as anchors for the IRT calibration. Because the items had an open character, we expected a student to work for approximately 35–40 min on the 10 items for each test session, as tested in our pilot project (Smit and Birri 2014). All test items were rated based on a detailed manual with four competence levels according to our rubric. Satisfying an inter-rater reliability of $Kappa > 0.70$ was expected for each of the rating teams (8 9 2 persons).

Missing data

Among the students who completed the questionnaire on their perceptions of feedback and instruction quality, self-concept, and motivation, individual missing values occurred. Missing values also occurred for a few teachers. Given that these missing values were not due to the design of the study, we assumed that they occurred randomly and consequently applied the full information maximum likelihood (FIML) procedure as a model-based treatment of missing data (Enders 2010).

Results

Teacher perceptions and effects of the intervention

To find an answer to the first hypothesis regarding the effects of the teachers' use of rubrics on their self-reported diagnostic skills and formative feedback practices, we developed a longitudinal model that included four variables: diagnostic skills, peer- and self-assessment, and formative feedback. In addition, we added transmissive beliefs of learning as a possible control variable. In general, the four variables showed higher values after the intervention, especially for peer- and self-assessment. The stability within the teacher sample with respect to the variables was high, with correlations between the two times of measurement ranging from 0.45 to 0.75. The results show that except for transmissive beliefs, the variables correlate (see Table 2) more strongly for T2 than for T1. Teacher beliefs did not change appreciably as part of the training, in line with other research (Swan 2007).

Because the competence of solving word problems might be affected by reading competence, we next analysed three longitudinal regression models for diagnostic skills, peer- and self-assessment, and formative feedback and included control variables. As control variables, we used the number of students with special needs and the number with foreign language background. We also added a dummy variable for the intervention/control group and the teachers' transmissive beliefs of learning mathematics. Only one control variable reached significance and only for the model with diagnostic skills: The number of students with a foreign language background has a significant negative affect on

Table 2 Descriptive statistics and correlations of teacher perceptions

	Mean	SD	1	2	3	4	5	6	7
1. Diagnostic skills T1	4.33	0.77							
2. Peer- and self-assessment T1	2.87	0.88	0.23						
3. Formative feedback T1	3.90	0.57	0.27	0.28					
4. Transmissive beliefs of learning math T1	3.20	0.77	0.16	-0.13	0.22				
5. Diagnostic skills T2	4.37	0.76	0.62**	0.16	0.37*	0.14			
6. Peer- and self-assessment T2	3.41	1.28	0.26	0.75**	0.43**	-0.03	0.19		
7. Formative feedback T2	4.12	0.80	0.40**	0.33*	0.45**	0.07	0.50**	0.32*	
8. Transmissive beliefs of learning math T2	3.22	0.84	0.06	-0.14	0.14	0.58**	0.10	-0.13	-0.02

$N = 45/43$; ** $p \leq 0.01$; * $p \leq 0.05$. Likert scale 1–6; poles: never/always or fully not agree/fully agree

the teachers' diagnostic skills ($b = -0.21$, $p = 0.056$, $p \leq 0.10$). As expected, transmissive beliefs do not explain any variance within the three models.

Based on these results and the theoretical model, we analysed the relationships between the three scales in a partly cross-lagged model (see Fig. 2). From a theoretical point of view, we expected effects from our intervention on diagnostic skills, formative feedback,

Fig. 2 Standardised coefficients for model of teachers' self-perceptions related to the intervention with rubrics. Bayesian estimates. $N = 44$; * $p \leq 0.05$

and self- and peer-assessment, whereas diagnostic skills is a predictor for formative feedback, as well. The model was estimated with aggregated manifest variable indicators (parcels of items) to reduce the number of parameters calculated in complex models owing to the sample size (Boivard and Koziol 2012). The model also contains the variable number of students with a foreign language background as a control variable for diagnostic skills.

We first calculated the model using MPlus 7 with maximum likelihood estimation ($CFI = 0.93$, $TLI = 0.91$, and $RMSEA = 0.09$). As the sample size is rather small, we switched to Bayesian estimation, which allows for better model estimation because large-sample theory is not needed (Muthén and Asparouhov 2012). In calculating the cross-lagged model in Fig. 2, Bayesian estimation with a Gibbs-Algorithm was applied. After conducting estimations for different iterations to determine convergence and PSR values, the outputs of the final model analysis produced stable results. The PPP value amounted to 0.30 and to an f difference of 15.33. The number of free parameters was 26; the deviance (DIC) was 579.25; the Bayesian (BIC) was 630.97, and the estimated number of parameters was 14.72. Plots for controlling ‘burn in’ and posterior distributions were satisfactory.

In the partly cross-lagged model (Fig. 2), diagnostic skills and peer- and self-assessment show strong autoregressive effects, and formative feedback has a medium stable effect. There is one cross-lagged effect only: Teachers with higher diagnostic skills at T1 appear to give more formative feedback after the intervention at T2. Two control variables affected the longitudinal relationship of diagnostic skills before and after the intervention. Teachers in the intervention group perceive their diagnostic skills after the intervention as being significantly higher than those in the control group ($b = 0.23$). In addition, the greater the number of students with a foreign language background, the less competent the teachers feel in diagnosing students’ problems working on tasks for mathematical reasoning ($b = -0.19$).

Student perceptions and effects of the intervention

The second and the third hypothesis relate to the effects of the use of rubrics on the formative assessment practices perceived by the students and their perceived competence for self-regulation and self-efficacy. We analysed four scales based on items from the student questionnaire and the results of the pre- and post-mathematical reasoning tests. In Table 3, the means of both time points and the correlations are presented. Significant increases are observed for all scales over time, with one significant decrease (formative feedback). Perceived practices with peer- and self-assessment, formative feedback, and perceived self-regulation competence significantly correlate with each other and with time. There is also a stable correlation between self-regulation and self-efficacy for mathematical reasoning ($r = 0.40$). With respect to the test-measurements, the relationship with other constructs is less evident; only self-efficacy shows a stronger coefficient at both time points ($r = 0.23/0.26$). These results are all in agreement with the research model and indicate that direct effects of our intervention on student outcomes might be difficult to prove.

In line with our research model, we constructed a longitudinal SEM (Fig. 3), with peer- and self-assessment and formative feedback being predictors for self-regulation. Because we applied a pre-/post-research design, both measurements of time were included in the model. To take into account the nested structure of the data—students in classes—we applied a multi-level SEM in which the covariance matrix was decomposed within and between levels (Hox 2010). Conceptually, relations at the classroom level consist of shared perceptions of students in one classroom, whereas relations at the student level pertain to

Table 3 Descriptive statistics and correlations of student perceptions and mathematical reasoning competence

	Mean	SD	ICC	1	2	3	4	5	6	7	8	9
1. Peer- and self-assessment T1	2.83	0.95	0.11									
2. Formative feedback T1	3.78	0.87	0.15	0.52**								
3. Self-regulation T1	3.70	0.76	0.13	0.31**	0.35**							
4. Self-efficacy reasoning T1	3.57	1.01	0.04	0.24**	0.19**	0.34**						
5. Mathematical reasoning T1	−0.02	0.82	0.14	−0.04	0.03	0.01	0.19**					
6. Peer- and self-assessment T2	3.02	1.01	0.21	0.31**	0.29**	0.20**	0.14**	0.08*				
7. Formative feedback T2	3.43	0.86	0.12	0.29**	0.48*	0.18**	0.11**	0.06	0.53**			
8. Self-regulation T2	3.91	0.81	0.12	0.18**	0.22**	0.47**	0.24**	0.07	0.34**	0.35**		
9. Self-efficacy reasoning T2	3.68	1.04	0.07	0.11*	0.09**	0.26**	0.53**	0.25**	0.20**	0.17**	0.37**	
10. Mathematical reasoning T2	0.41	0.81	0.16	−0.03	0.03	0.03	0.22**	0.96**	0.08*	0.04	0.08*	0.26**

$N = 762$; ** $p < 0.01$, * $p < 0.05$. Likert scale 1–6; poles: never/always or fully not agree/fully agree

Fig. 3 Standardised coefficients for model of students' perceptions related to self-regulation and the intervention with rubrics. $N_{\text{between}} = 44$, $N_{\text{within}} = 762$; * $p \leq 0.05$

individual differences between students within one classroom. A first model accordingly showed insufficient fit indices. Based on the analysis of the path coefficients and the modification indices suggested by Mplus, we made some small adjustments at the class level. The final model (Fig. 3) established for testing effects on self-regulation showed the following good fit indices: CFI = 0.98, TLI = 0.96, RMSEA = 0.04, SRMR_{within} = 0.03, and SRMR_{between} = 0.20. SRMR_{between} is rather large, which might be due to the small ICC and small sample size at the between level (Hsu et al. 2016). Therefore, we reanalysed our model with Bayesian estimation, which is less sample size dependent for misspecifications. Path coefficients in the Bayesian model showed coefficients comparable to those of the default ML-estimator.

The within part of the model shows medium stable paths for self-regulation and formative feedback, whereas the coefficient for peer- and self-assessment ($b = 0.19$) indicates less stability. Self-regulation as perceived by the students is influenced by peer- and self-assessment and formative feedback. There is a cross-lagged effect over time of formative feedback on peer- and self-assessment, as theoretically expected. At the between level, all autoregressive paths are of medium or high stability. Our intervention has a significant effect on peer- and self-assessment ($b = 0.61$) and on formative feedback ($b = 0.38$) as perceived by the class. We could not include the cross-lagged effect as we did at the within level. The model explains $R^2 = 0.79$ of the variance of self-regulation at the between level. Self-regulation is predicted by peer- and self-assessment and related to formative feedback. Thus, our intervention has a mediated effect on self-regulation, as well.

We applied the same model for self-efficacy as for self-regulation (Fig. 4). The good fit indices are comparable: CFI = 0.98, TLI = 0.96, RMSEA = 0.04, SRMR_{within} = 0.03, and SRMR_{between} = 0.28. Overall, the coefficients are comparable. The intervention has a

Fig. 4 Standardised coefficients for model of students' perceptions related to self-efficacy and the intervention with rubrics. $N_{\text{between}} = 44$, $N_{\text{within}} = 762$; * $p \leq 0.05$

comparable significant effect on peer- and self-assessment ($b = 0.64$) and on formative feedback ($b = 0.34$) as perceived by the class. In addition, there is a mediated effect of the intervention on self-efficacy. The model explains $R^2 = 0.84$ of the variance of self-efficacy at the class level.

With the fourth hypothesis, we explored the effects of the rubrics on the student outcomes and which indirect effects might play a role. Therefore, we relate to the research model (Fig. 1), in which an indirect effect of the use of rubrics mediated by self-regulation and self-efficacy is suggested. In line with this theoretical model, we established a number of multi-level regression models to demonstrate the effects of the intervention, perceived self-regulation competence, and self-efficacy. Several control variables, such as gender, student nationality, and age, were tested for bias effects, but none were significant. Prior test performance, a grouping variable, and the two predictors were added in a stepwise fashion in the following regression models (Table 4, Models 1–4). Reasoning test performance T1 was a significant predictor of post-test scores (Model 1) and explains almost all variance on both levels (individual/class). There is no direct effect of the intervention on reasoning competence T2 (Model 2). However, the grouping variable adds 1% to the explained between variance. Self-efficacy is a significant but negligible predictor at the individual level (Model 3), becoming slightly stronger in Model 4, in which self-regulation is added. Self-regulation is not significant at either level, although the coefficient at the class level reaches $b = 0.14$.

Table 4 Multilevel regression analyses predicting students' mathematical reasoning competence after the intervention (T2)

	Model 1		Model 2		Model 3		Model 4	
	b	SE	b	SE	b	SE	b	SE
Individual-level variables								
Mathematical reasoning T1	0.95*	0.00	0.95*	0.00	0.94*	0.00	0.94*	0.00
Self-efficacy reasoning T2					0.03*	0.01	0.04*	0.01
Self-regulation T2							-0.01	0.01
Class-level variables								
Mathematical reasoning T1	0.97*	0.01	0.96*	0.02	0.95*	0.03	0.99*	0.02
Intervention group (=1)			0.04	0.05	0.05	0.07	0.03	0.07
Self-efficacy reasoning T2					-0.04	0.09	-0.13	0.11
Self-regulation T2							0.14	0.10
R^2 within	0.91*	0.00	0.91*	0.00	0.91*	0.00	0.91*	0.00
R^2 between	0.93*	0.03	0.94*	0.03	0.92*	0.04	0.94*	0.04

SE standard error

$N_{between} = 44$, $N_{within} = 762$

* $p < 0.05$

Discussion

This article presents results from a study aimed to test the effects of a formative rubric within a complex learning environment. As suggested by Brookhart and Chen (2015), we used an experimental control-group design for our research and avoided confounding of rubrics with other aspects of treatment (feedback, self-assessment, and so on) by using an identical lesson script for both groups. Despite this controlled procedure, our research is of high ecological validity and thus usable in a practical sense (Sandoval and Bell 2004).

In general, the theoretical model of Panadero and Jonsson (2013) about rubrics and the moderating factors on students' outcomes proved to be a useful basis for our SE models. With respect to Hypothesis (1a), we could show that our intervention had a positive effect on the self-reported teacher's diagnostic skills but not on the frequency of formative feedback. However, prior diagnostic skills affect later formative feedback. Thus, our rubric has an indirect effect on feedback. As predicted in Hypothesis (1b), our intervention had no effect on the frequency of the teachers' self-perceived peer- and self-assessment practices. In both groups, teachers were expected to conduct a certain number of peer- and self-assessment practices.

However, based on the student data, there is a direct effect of our intervention on formative feedback as well as on peer- and self-assessment practices. Therefore, we can confirm Hypothesis (2). Why is there a difference in the teacher and the student effects? Hattie (2008) suggests that teachers often claim they give feedback, but typically they do not. When we confront teacher perceptions with those of students, it shows that there is a low correlation for peer- and self-assessment ($r = 0.17$) and a high correlation for formative feedback ($r = 0.40$). The data triangulation suggests that at least for peer- and self-assessment, the teacher might have a biased perception. There is an indirect effect of our intervention on the students' self-regulation and self-efficacy via peer- and self-assessment and related to formative feedback (Hypothesis 3). Hypothesis (4) was treated in a more

explorative way. The intervention showed no direct effect of the rubric on the students' reasoning competence. There are significant correlations between self-efficacy and reasoning competence at both time points. However, proving effects is difficult. Regarding mediated effects of self-regulation or self-efficacy, it appears that almost all variance of the outcome variable is explained by prior reasoning competence. Hence, other predictors do not add much and are not very strong. Only self-efficacy at the individual level indicates that the theoretically expected causal links as shown in Fig. 1 could exist. Therefore, our model should be adapted and show self-regulation competence and reasoning competence as two independent student outcomes linked by self-efficacy. This is also confirmed by the fact that there is a significant correlation between items asking for the students' perceived utility of the rubric within the intervention group at T2 and students' self-efficacy and self-regulation competence. There is no correlation of perceived utility of the rubric and reasoning competence, however. In the intervention group, not only the competent students but also students at different levels of competence in reasoning evaluated the rubric as helpful for making the goals more clear, providing security or motivating (perceived utility). The missing direct effect of the use of rubrics on student achievement in our study is in agreement with the theoretical model (Fig. 1). Even if, as we could show, rubrics only indirectly have an impact on achievement, 'the fact that rubrics are an efficient, clear, and easily understood way to focus learning goals, criteria, and performance descriptions would recommend their use' (Brookhart and Chen 2015, p. 363). Our results are partly in line with those from the research by Panadero et al. (2012): In their study, students at the end of secondary school using a rubric in geography showed higher self-regulation competencies than those in the control group. They also developed higher self-efficacy beliefs in combination with provided process feedback in the rubric group. Contrary to our results, the rubric group performed better than the control group, which was also the case in the study on rubrics for writing essays by Andrade et al. (2008). Following a line of argumentation by Ross et al. (2002), the missing clearer effect on reasoning might be due to the nature of mathematics and teachers' representation of it. For some teachers, teaching mathematics related to problem-solving could have been rather difficult, more complex than, e.g., teaching language. Thus, in line with Ross et al. (2002), teacher training about alternative assessment methods should be included for the teachers in need of coaching on subject knowledge during the implementation.

In our study, there is an unexpected negative development of feedback frequency and quality in both groups as perceived by the students over time. From their perspective, it looks like our teachers provided less feedback during the intervention than they normally did in exchange for letting the students do more self- and peer-assessment. This is important because, to acquire self-assessment competence, students need opportunities to assess themselves and each other. Peer assessment help develop the competence needed to make objective judgements against standards (Nicol and Macfarlane-Dick 2006). Some students might have shown more growth if the teacher had given more individual feedback as part of scaffolding to students in need. This result seems contradictory at least for the rubric group, where we expected the teacher to offer rich feedback. The teachers even stated to have given more formative feedback on T2. One explanation could be that providing feedback that is specific and clear, for conceptual and procedural learning tasks, is a reasonable, general guideline. Mathematical reasoning tasks do not allow for simple true/false responses, and feedback even on rubrics might sometimes appear more unclear or even ambiguous to the students compared with feedback for arithmetic calculations. We know from research that formative feedback effects are stronger for memory tasks and weaker for more procedural tasks (Kluger and DeNisi 1996).

The effects of our study for teacher diagnosing and the use of peer- and self-assessment with the help of a rubric are clearer than for formative feedback. Therefore, further research should examine, e.g., from an observer's perspective, the feedback quality based on rubrics.

A limitation of our study, especially with respect to the calculation of the standard errors in the complex SEM models, is the low sample size on the second level. As a consequence, the results obtained from mediators related to student achievement might not become significant until a larger experimental group is considered. Although we targeted a sufficiently large sample size of at least 50 teachers (to reach a power of 0.8), we had difficulties motivating teachers to participate voluntarily in our extensive development program. As a next step, we intend to explore the sustainability of our project by analysing our third wave of measurement.

Acknowledgements The presented project was funded by the Swiss National Science Foundation (project no. 149386).

References

- Andrade, H. (2000). Using rubrics to promote thinking and learning. *Educational Leadership*, 57(5), 13–18.
- Andrade, H., Du, Y., & Mycek, K. (2010). Rubric-referenced self-assessment and middle school students' writing. *Assessment in Education: Principles, Policy & Practice*, 17(2), 199–214.
- Andrade, H., Du, Y., & Wang, X. (2008). *Putting rubrics to the test: The effect of a model, criteria generation, and rubric-referenced self-assessment on elementary school students' writing* (pp. 3–13). Educational Measurement: Issues and Practice.
- Andrade, H., Wang, X., Du, Y., & Akawi, R. L. (2009). Rubric-referenced self-assessment and self-efficacy for writing. *The Journal of Educational Research*, 102(4), 287–302. doi:10.3200/JOER.102.4.287-302.
- Arter, J. A., & Chappuis, J. (2007). *Creating & recognizing quality rubrics*. Upper Saddle River, NJ: Pearson Education.
- Ash, D., & Levitt, K. (2003). Working within the zone of proximal development: Formative assessment as professional development. *Journal of Science Teacher Education*, 14(1), 23–48.
- Bandura, A. (1993). Perceived self-efficacy in cognitive development and functioning. *Educational psychologist*, 28(2), 117–148.
- Bangert-Drowns, R. L., Kulik, C.-L. C., Kulik, J. A., & Morgan, M. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research*, 61(2), 213–238.
- Bezold, A. (2009). *Förderung von Argumentationskompetenzen durch selbstdifferenzierende Lernangebote. Eine Studie im Mathematikunterricht der Grundschule* [Fostering argumentation competence through self-differentiating learning environments. A study in mathematics education in elementary school]. Hamburg: Kovac.
- Blanton, M. L., & Kaput, J. J. (2005). Characterizing a classroom practice that promotes algebraic reasoning. *Journal for Research in Mathematics Education*, 36(5), 412–446.
- Blum, W., & Kirsch, A. (1991). Preformal proving: Examples and reflections. *Educational Studies in Mathematics*, 22(2), 183–203. doi:10.2307/3482408.
- Boivard, J. A., & Koziol, N. A. (2012). Measurement models for ordered-categorical indicators. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 495–511). New York, NY: The Guildford Press.
- Bransford, J. D., & Donovan, M. S. (Eds.). (2005). *How students learn: History, mathematics, and science in the classroom*. Washington, DC: National Research Council (U.S.), Committee on How People Learn, A Targeted Report for Teachers.
- Brodie, K. (2010). *Teaching mathematical reasoning in secondary school classrooms*. New York: Springer.
- Brookhart, S. M. (2013). *How to create and use rubrics for formative assessment and grading*. Alexandria, VA: Ascd.
- Brookhart, S. M., & Chen, F. (2015). The quality and effectiveness of descriptive rubrics. *Educational Review*, 67(3), 343–368. doi:10.1080/00131911.2014.929565.
- Brown, G. T. L., Harris, L. R., & Harnett, J. (2012). Teacher beliefs about feedback within an assessment for learning environment: Endorsement of improved learning over student well-being. *Teaching and Teacher Education*, 28(7), 968–978. doi:10.1016/j.tate.2012.05.003.
- Brunner, M., Anders, Y., Hachfeld, A., & Krauss, S. (2013). The diagnostic skills of mathematics teachers. In M. Kunter, J. Baumert, W. Blum, U. Klusmann, S. Krauss, & M. Neubrand (Eds.), *Cognitive activation in the mathematics classroom and professional competence of teachers: Results from the COACTIV Project* (pp. 229–248). Boston: Springer.
- Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research*, 65(3), 245–281.
- Christou, C., & Papageorgiou, E. (2007). A framework of mathematics inductive reasoning. *Learning and Instruction*, 17(1), 55–66. doi:10.1016/j.learninstruc.2006.11.009.
- de Lange, J. (1999). *Framework for classroom assessment in mathematics*. Utrecht/Madison, WI: Freudenthal Institute/University of Madison, WI, National Centre for Improving Student Learning and Achievement in Mathematics and Science.
- Earl, L. M. (2012). *Assessment as learning: Using classroom assessment to maximize student learning*. Thousand Oaks, CA: Corwin Press.
- Enders, C. K. (2010). *Applied missing data analysis*. New York: Guildford.
- Ginsburg, H. P. (2009). The challenge of formative assessment in mathematics education: Children's minds, teachers' minds. *Human Development*, 52, 109–128.
- Hamers, J. H. M., de Koning, E., & Sijtsma, K. (1998). Inductive reasoning in third grade: Intervention promises and constraints. *Contemporary Educational Psychology*, 23(2), 132–148. doi:10.1006/ceps.1998.0966.
- Hargreaves, E., McCallum, B., & Gipps, C. (2000). Teacher feedback strategies in primary classrooms—new evidence. In A. Susan (Ed.), *Feedback for learning* (pp. 21–31). London: Routledge.

- Hattie, J. (2008). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. London: Routledge.
- Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1), 81–112.
- Hox, J. J. (2010). *Multilevel analysis: Techniques and applications*. New York: Routledge.
- Hsia, L.-H., Huang, I., & Hwang, G.-J. (2015). A web-based peer-assessment approach to improving junior high school students' performance, self-efficacy and motivation in performing arts courses. *British Journal of Educational Technology*, n/a-n/a. doi:10.1111/bjet.12248.
- Hsu, H.-Y., Lin, J.-H., Kwok, O.-M., Acosta, S., & Willson, V. (2016). The impact of intraclass correlation on the effectiveness of level-specific fit indices in multilevel structural equation modeling: A Monte Carlo study. *Educational and Psychological Measurement*. doi:10.1177/0013164416642823.
- Keiser, J. M., & Lambdin, D. V. (1996). The clock is ticking: Time constraint issues in mathematics teaching reform. *The Journal of Educational Research*, 90(1), 23–31. doi:10.1080/00220671.1996.9944440.
- Kempert, S., Saalbach, H., & Hardy, I. (2011). Cognitive benefits and costs of bilingualism in elementary school students: The case of mathematical word problems. *Journal of Educational Psychology*, 103(3), 547.
- Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, 119(2), 254–284.
- Kroesbergen, E. H., & Van Luit, J. E. (2003). Mathematics interventions for children with special educational needs a meta-analysis. *Remedial and Special Education*, 24(2), 97–114.
- Libbee, M. (2001). Assessment as a diagnostic tool. *Journal of Geography*, 100(4), 175–178. doi:10.1080/00221340108978437.
- Mui So, W. W., & Hoi Lee, T. T. (2011). Influence of teachers' perceptions of teaching and learning on the implementation of Assessment for Learning in inquiry study. *Assessment in Education: Principles, Policy & Practice*, 18(4), 417–432. doi:10.1080/0969594X.2011.577409.
- Mullis, I. V., Martin, M. O., Gonzalez, E. J., & Chrostowski, S. J. (2004). *TIMSS 2003 international mathematics report: Findings from IEA's trends in international mathematics and science study at the fourth and eighth grades*. Boston: IEA International Association for Evaluation of Educational Achievement.
- Muthén, B. O., & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological Methods*, 17(3), 313–335. doi:10.1037/a0026802.
- National Council of Teachers of Mathematics (NCTM). (2000). *Principles and standards for school mathematics*. Reston VA: National Council of Teachers of Mathematics.
- Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education*, 31(2), 199–218. doi:10.1080/03075070600572090.
- Organization for Economic Co-Operation and Development (OECD). (2004). *Problem solving for tomorrow's world. First measures of cross-curricular competencies from PISA 2003*. Paris: OECD.
- Organization for Economic Co-Operation and Development (OECD). (2014). *PISA 2012 results: What students know and can do—student performance in mathematics, reading and science*. Paris: PISA, OECD.
- Pajares, F., & Graham, L. (1999). Self-efficacy, motivation constructs, and mathematics performance of entering middle school students. *Contemporary Educational Psychology*, 24(2), 124–139. doi:10.1006/ceps.1998.0991.
- Panadero, E., & Jonsson, A. (2013). The use of scoring rubrics for formative assessment purposes revisited: A review. *Educational Research Review*, 9, 129–144. doi:10.1016/j.edurev.2013.01.002.
- Panadero, E., Jonsson, A., & Srijbos, J.-W. (2016). Scaffolding self-regulated learning through self-assessment and peer assessment: Guidelines for classroom implementation. In D. Laveault & L. Allal (Eds.), *Assessment for learning: Meeting the challenge of implementation* (pp. 311–326). Cham: Springer International Publishing.
- Panadero, E., Tapia, J. A., & Huertas, J. A. (2012). Rubrics and self-assessment scripts effects on self-regulation, learning and self-efficacy in secondary education. *Learning and Individual Differences*, 22(6), 806–813. doi:10.1016/j.lindif.2012.04.007.
- Perry, N. E., Phillips, L., & Dowler, J. (2004). Examining features of tasks and their potential to promote self-regulated learning. *The Teachers College Record*, 106(9), 1854–1878.
- Perry, N. E., & Winne, P. H. (2006). Learning from learning kits: gStudy traces of students' self-regulated engagements with computerized content. *Educational Psychology Review*, 18(3), 211–228.
- Phye, G. D., & Sanders, C. E. (1994). Advice and feedback: Elements of practice for problem solving. *Contemporary Educational Psychology*, 19(3), 286–301. doi:10.1006/ceps.1994.1022.
- Popham, W. J. (1997). What's wrong—and what's right—with rubrics. *Educational Leadership*, 10, 72–75.
- Purdie, N., Hattie, J., & Douglas, G. (1996). Student conceptions of learning and their use of self-regulated

- learning strategies: A cross-cultural comparison. *Journal of Educational Psychology*, 88(1), 87.
- Rakoczy, K., Buff, A., Lipowsky, F., Hugener, I., Pauli, C., & Reusser, K. (2005). *Dokumentation der Erhebungs- und Auswertungsinstrumente zur schweizerisch-deutschen Videostudie "Unterrichtsqualität, Lernverhalten und mathematisches Verständnis"* [Documentation of the survey instruments in the Swiss-German video study "Teaching quality, learning behaviour and mathematical comprehension"]. Frankfurt am Main: Gesellschaft zur Förderung Pädagogischer Forschung.
- Ross, J. A., Hogaboam-Gray, A., & Rolheiser, C. (2002). Student self-evaluation in grade 5–6 mathematics effects on problem-solving achievement. *Educational Assessment*, 8(1), 43–58. doi:10.1207/S15326977EA0801_03.
- Saddler, B., & Andrade, H. (2004). The writing rubric. *Educational Leadership*, 62(2), 48–52.
- Sandoval, W. A., & Bell, P. (2004). Design-based research methods for studying learning in context: Introduction. *Educational Psychologist*, 39(4), 199–201.
- Semadeni, Z. (1984). Action proofs in primary mathematics teaching and in teacher training. *For the Learning of Mathematics*, 4(1), 32–34.
- Sfard, A. (2001). There is more to discourse than meets the ears: Looking at thinking as communicating to learn more about mathematical learning. *Educational Studies in Mathematics*, 46(1/3), 13–57.
- Shute, V. J. (2008). Focus on formative feedback. *Review of educational research*, 78(1), 153–189.
- Smit, R. (2009). *Die formative Beurteilung und ihr Nutzen für die Entwicklung von Lernkompetenz [Formative assessment and the benefits for the development of learning competence]*. Baltmannsweiler: Schneider Verlag Hohengehren.
- Smit, R., & Birri, T. (2014). Assuring the quality of standards-oriented classroom assessment with rubrics for complex competencies. *Studies in Educational Evaluation*, 43, 5–13. doi:10.1016/j.stueduc.2014.02.002.
- Stanat, P., & Christensen, G. (2006). *Where immigrant students succeed: A comparative review of performance and engagement in PISA 2003*. Paris: OECD.
- Stein, M. K., Grover, B. W., & Henningsen, M. (1996). Building student capacity for mathematical thinking and reasoning: An analysis of mathematical tasks used in reform classrooms. *American Educational Research Journal*, 33(2), 455–488.
- Swan, M. (2007). The impact of task-based professional development on teachers' practices and beliefs: A design research study. *Journal of Mathematics Teacher Education*, 10(4–6), 217–237.
- Swanson, H. L. (2014). Does cognitive strategy training on word problems compensate for working memory capacity in children with math difficulties? *Journal of Educational Psychology*, 106(3), 831–848. doi:10.1037/a0035838.
- Swiss Conference of Cantonal Ministers of Education (EDK). (2011). *Nationale Bildungsziele für die obligatorische Schule: in vier Fächern zu erreichende Grundkompetenzen* [National standards for compulsory education: Basic competences for four subjects]. Retrieved January 10, 2012, from EDK, http://www.edudoc.ch/static/web/arbeiten/harmos/grundkomp_faktenblatt_d.pdf.
- The Standing Conference of the Ministers of Education and Cultural Affairs of the Länder in the Federal Republic of Germany (KMK). (2004). *Bildungsstandards im Fach Mathematik für den Hauptschulabschluss nach Klasse 9* [Standards for the subject mathematics in class 9 end of junior high school]. http://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2004/2004_10_15-Bildungsstandards-Mathe-Primar.pdf.
- Thompson, D. R., & Schultz-Ferrel, K. (2008). *Introduction to reasoning and proof: Grades 6–8*. Portsmouth, NH: Heinemann.
- Turner, S. L. (2014). Creating an assessment-centered classroom: Five essential assessment strategies to support middle grades student learning and achievement. *Middle School Journal*, 45(5), 3–16.
- Underwood, J. S., & Tregidgo, A. P. (2006). Improving student writing through effective feedback: Best practices and recommendations. *Journal of Teaching Writing*, 22(2), 73–98.
- Voss, T., Kleickmann, T., Kunter, M., & Hachfeld, A. (2013). Mathematics teachers' beliefs. In M. Kunter, J. Baumert, W. Blum, U. Klusmann, S. Krass, & M. Neubrand (Eds.), *Cognitive activation in the mathematics classroom and professional competence of teachers* (pp. 249–271). New York: Springer.
- Wollenschläger, M., Hattie, J., Machts, N., Möller, J., & Harms, U. (2016). What makes rubrics effective in teacher-feedback? Transparency of learning goals is not enough. *Contemporary Educational Psychology*, 44–45, 1–11. doi:10.1016/j.cedpsych.2015.11.003.
- Zimmerman, B. J. (2000). Self-efficacy: An essential motive to learn. *Contemporary Educational Psychology*, 25(1), 82–91.
- Zimmerman, B. J. (2008). Investigating self-regulation and motivation: Historical background, methodological developments, and future prospects. *American Educational Research Journal*, 45(1), 166–183.