Research Article https://doi.org/10.12973/eu-jer.11.2.663

European Journal of Educational Research
Volume 11, Issue 2, 663 - 680.
ISSN: 2165-8714
https://www.eu-jer.com/

Observed Quality of Formative Peer and Self-Assessment in Everyday Mathematics Teaching and its Effects on Student Performance

Sandra Zulliger*
University of Teacher Education Lucerne, SWITZERLAND

Alois Buholzer
University of Teacher Education Lucerne, SWITZERLAND

Merle Ruelmann
University of Teacher Education Lucerne, SWITZERLAND

Received: November 17, 2021 ▪ Revised: December 26, 2021 ▪ Accepted: January 6, 2022

Abstract: The positive effect of peer assessment and self-assessment strategies on learners' performance has been widely confirmed in experimental or quasi-experimental studies. However, whether peer and self-assessment within everyday mathematics teaching affect student learning and achievement, has rarely been studied. This study aimed to determine with what quality peer and self-assessment occur in everyday mathematics instruction and whether and which students benefit from it in terms of achievement and the learning process. Two lessons on division were video-recorded and rated to determine the quality of peer and self-assessment. Six hundred thirty-four students of fourth-grade primary school classes in German-speaking Switzerland participated in the study and completed a performance test on division. Multilevel analyses showed no general effect of the quality of peer or self-assessment on performance. However, high-quality self-assessment was beneficial for lower-performing students, who used a larger repertoire of calculation strategies, which helped them perform better. In conclusion, peer and self-assessment in real-life settings only have a small effect on the student performance in this Swiss study.

Keywords: *Everyday mathematics teaching, formative assessment, learning process, peer assessment, self-assessment.*

To cite this article: Zulliger, S., Buholzer, A., & Ruelmann, M. (2022). Observed quality of formative peer and self-assessment in everyday mathematics teaching and its effects on student performance. *European Journal of Educational Research, 11*(2), 663-680. https://doi.org/10.12973/eu-jer.11.2.663

Introduction

Peer assessment and self-assessment (PASA) are considered two of the five key elements of formative assessment (Black & Wiliam, 2009). Formative assessment is used to collect diagnostic information about learning and its outcomes in the classroom, in order to use it to improve teaching and learning processes (Schütze et al., 2018). In addition to summative assessment, which measures accumulated learning, formative assessment is carried out during the ongoing learning process (Cizek, 2010). While peer assessment (PA) focuses on collaborative and cooperative reflection, assessment, and sending and receiving peer feedback (Kollar & Fischer, 2010; Strijbos & Wichmann, 2017), self-assessment (SA) focuses on self-reflection and assessment with self-feedback (Andrade & Valtcheva, 2009). Both approaches emphasise student feedback on themselves or peers, and increase student involvement in assessment, stimulating metacognitive processes and optimising individual learning (Black & Wiliam, 2009; Panadero, 2016). Several meta-studies have demonstrated the positive effect of PASA on students' learning performance (Graham et al., 2015; Sanchez et al., 2017). These previous studies are often based on (quasi-)experimental designs or self-reports. However, such studies may have low ecological validity, due to a lack of transferability to the everyday context (Schnell et al., 2013). There is still little research on PASA in "real learning settings" (Panadero et al., 2016, p. 811), despite its high potential for enhancing learning. Furthermore, it remains unclear whether all learners benefit equally from PASA and why and how it affects the learning process and the outcome (Andrade, 2019). Empirical evidence suggests that lower-performing learners could benefit from PASA in terms of the learning process and the associated performance (Gersten et al., 2009; Kingston & Nash, 2011). The current paper addresses these research gaps by analysing the observed quality of PASA in everyday mathematics lessons with regard to the learning process and outcome on students of different ability levels. The research questions are: What quality of PASA occurs in everyday (non-experimental) teaching settings? Does PASA quality positively affect student performance or metacognitive activity, operationalised as using different calculation strategies? Do students of different ability levels benefit equally from PASA quality?

* **Corresponding author:**
Sandra Zulliger, Institute for Diversity in Education, University of Teacher Education Lucerne, Lucerne, Switzerland. ✉ sandra.zulliger@phlu.ch

A study in a real-life setting was conducted with 52 Swiss school classes in order to answer the research questions. Two mathematics lessons were video-recorded from the 52 school classes in the fourth primary level and rated in terms of PASA quality. The 634 participating students completed a mathematics performance test at the beginning and end of the division topic. The performance test was used to determine the increase in performance and the choice of different calculation strategies. The variability of calculation strategies was chosen as an indicator of metacognitive activity. External, trained observers assess the quality of PASA. A high level of implementation quality of PASA can be observed in challenging tasks, in careful guidance and monitoring by the teacher, and use of the information gained for further teaching and learning processes (Andrade & Valtcheva, 2009; Topping, 2009; Wylie & Lyon, 2013).

Literature Review

Defining Quality of Formative PASA

Both PA and SA can support the learning process if they provide information that teachers and their students can use as feedback to assess themselves and modify teaching and learning activities. Formative PASAs are successful when the students' metacognitive processes are triggered and enhance learning (Black et al., 2004). According to Topping (2009), PA can be defined as "an arrangement for learners to consider and specify the level, value, or quality of a product or performance of other equal-status learners" (p. 20).

Formative SA is described by Andrade (2010) as "a process of formative assessment during which students reflect on the quality of their work, judge the degree to which it reflects explicitly stated goals or criteria, and revise their work accordingly" (p. 91). SA can be seen as feedback to oneself (Andrade, 2010). Learners engage in a metacognitive dialogue by commenting on their learning outcomes, comparing them to the relevant standards, and drawing conclusions for further learning steps and revision of the work (Panadero et al., 2016). According to Wylie and Lyon (2013), this provides learners with the opportunity to reflect metacognitively on their learning process. From a theoretical perspective, the following three conditions for success can be identified for a learning-effective implementation of PASA: 1.) Level of challenge, 2.) Guidance, implementation, and support, 3.) Use of the information gained for further learning and teaching.

Level of Challenge: High-quality PASAs involve challenging tasks (Black et al., 2004; Wylie & Lyon, 2013), which challenge learners to engage in deeper thinking and metacognitive thought processes (Topping, 2009). These processes encourage students to describe and discuss their own or others' learning processes and results in a differentiated manner, and explicitly review and assess them, based on learning objectives and assessment criteria. PA should be elaborate and specific, lead to a dialogue between the participants (Kollar & Fischer, 2010), and relate supportively to learning and self-regulation (Alqassab, 2016). SA should be designed to help learners gain a deeper understanding of their own learning processes and support them in managing and further optimising their learning (Andrade & Valtcheva, 2009; Wylie & Lyon, 2013).

Guidance, Implementation, and Support: Teachers, especially at the primary school level, have the important task of carefully guiding and monitoring the implementation of PASA and offering them support in any necessary reflection and assessment processes (Harris & Brown, 2013; Ploegh et al., 2009). Specific instructions and careful management of interpersonal relationships are important for successfully implementing PASA (Harris & Brown, 2013). In addition, learners need to develop skills to apply PASA (Black et al., 2004).

Use of the Information Gained for Further Learning and Teaching: PASA becomes effective for learning when learners are encouraged to use the (metacognitive) insights from PASA to improve their learning processes. Effective use of the information gained means sharing, discussing, and systematising findings from the PASA in the classroom or guiding learners to consider the information gained from their own and others' feedback for further learning (Andrade & Valtcheva, 2009). For teachers, the challenge is to incorporate PASA into the structure of teaching and learning processes so that subsequent phases of instruction refer to it.

Measuring the Quality of PASA in Everyday Teaching

The quality of PASA in the classroom can be captured through self-evaluation or classroom observation, by establishing observable indicators and judging them (Pauli, 2012). Three different perspectives can be used for this purpose: external trained observers, teachers, and learners. Comparisons of perspectives reveal large differences between these groups of people (Fauth et al., 2014; Praetorius, 2014; Waldis et al., 2010). In contrast to judgments by teachers and students, external observers have the advantage that systematic comparisons across different classes and existing teaching theories are possible. Especially on didactic and methodological aspects of teaching, like the use and quality of PASA, are the evaluations of external trained observers considered a significant source of information due to their high validity in empirical teaching research (Waldis et al., 2010). Observational studies are also ecologically more valid than other research designs (Schnell et al., 2013), such as the (quasi-)experimental designs often used in PASA studies.

The few studies on assessments of everyday teaching by trained external observers are American, and conclude that the quality of PASA is rather low. Gotwals et al. (2015) found differences between teachers of different subjects, but none

reached the expert level. The results of the observational study by Oswalt (2013) point in the same direction, in which quality scores in the lower scale range were found for PASA. Neither Gotwals et al. (2015) nor Oswalt (2013) examined the relationship between PASA and student achievement.

Effects of PASA on Performance

There is a range of evidence from (meta-)studies on the effectiveness and mode of action of PASA on learner performance. In their meta-study, Double et al. (2019) demonstrate a small to medium-size effect on academic achievement for PA. Furthermore, the effects of PA are robust to contextual factors, and are evident at all school levels. The effects of SA are not significantly different from PA. The meta-study by Graham et al. (2015), with ten studies from grades 1 to 8, shows a significant positive effect on writing quality for PASA.

Sanchez et al.'s (2017) meta-study, including 33 studies with students from third to twelfth grade, indicates that learners performed better on subsequent tests after self-grading or peer-grading than learners without such grading. The meta-analysis of Brown and Harris (2013) found that the median effect of SA on academic performance lay between .40 and .45., but for some studies, the effects of SA were nil to small. Brown and Harris (2013) suggest that the different effects of the studies are due to qualitative differences in the implementation and to the cognitive demands of SA.

Some findings also indicate that the effectiveness of PASA depends on learner prerequisites. There is evidence that learners with weak school performance benefit primarily from PASA (Brown & Harris, 2013). For example, Gersten et al. (2009) state that low-performing learners benefit from formative assessment of their learning growth. This finding can also be found concerning learners with weak language skills at the primary school level (Brookhart et al., 2010; Decristan et al., 2015). Regarding PA, Gielen et al. (2010) showed that peer feedback improves writing performance, especially for learners who scored weakly on the pretest. According to Strijbos et al. (2010), a possible explanation for this finding is that the effectiveness of peer feedback depends on whether the peer is perceived as an expert and the student can really benefit accordingly.

Results on the differential impact of SA are mixed. Some studies suggest that low-achieving learners benefit from SA (Ross et al., 1999; Sadler & Good, 2006). On the other hand, Boud et al. (2013) showed that higher education students with average performance are more likely to gain accurate SA. Given these inconsistent research findings, Panadero et al. (2016) maintain that more studies should be conducted in "real learning settings" for a " better understanding of what happens in terms of cognitive, metacognitive, motivational, and emotional processes while students are self-assessing." Furthermore, "the differentiation of participants by their achievement levels (e.g., high vs low) would identify the effects in different kinds of students" (p. 811).

The positive impact of PASA on achievement is often explained in terms of promoting students' metacognitive skills (e.g., Pantiwati & Husamah, 2017). Several studies have shown that low performers can significantly improve their performance through metacognitive training (Cardelle-Elawar, 1995; Pennequin et al., 2010). Compared to high-achieving students, low performers benefit particularly from metacognitive training (Pennequin et al., 2010). Low-performing students usually lack the metacognition and self-monitoring that they need to know and measure their performance (Hill, 2016).

Metacognitive processes enable students to explicitly select and invent strategies by understanding the task requirements, their available cognitive resources, and their own experiences in solving similar problems (Pennequin et al., 2010). In division, the variable and adaptive use of calculation strategies are generally considered an important prerequisite for successful learning processes in mathematics didactics (Heinze et al., 2009). Therefore, high variability or a large repertoire of strategies per se is not considered a goal, but an important prerequisite for the selection and efficient use of strategies, for the learning process and later success in advanced arithmetic (Lindberg et al., 2013). Adaptive use of strategies is demonstrated when learners choose the strategy, which is optimally suited to the task, their abilities, and the context (e.g., criteria such as speed) (Hickendorff et al., 2019). When children use different strategies over a longer period, there is an overall trend toward using more advanced and adaptive strategies (Chen & Siegler, 2000).

The extent to which variable and adaptive numeracy skills can be promoted through instruction and, more narrowly, with formative PASA, has been little researched empirically (Heinze et al., 2009). Some initial indications can be found in Hartnett (2007). According to their results, the number of strategies in whole-number-based addition and subtraction can be increased in third graders, if they are regularly asked to show their thinking in the lessons.

Independent of the mathematical achievement level, learners use multiple strategies and adaptively apply them to arithmetic tasks with different difficulty levels (Lindberg et al., 2013; Siegler, 1996). However, a closer look reveals that learners with weak mathematics performance predominantly use ineffective, less cognitively demanding strategies (Geary & Hoard, 2005), and use them counter-adaptively (Fagginger Auer et al., 2016).

As mentioned before, particularly low-performing students benefit from metacognitive interventions (Hill, 2016; Pennequin et al., 2010). The presented research findings lead us to the following hypothesis:

H1a: SA quality and PA quality positively affect strategy variability for division problems (Geurten & Lemaire, 2019).
H1b: Particularly low-performing students (i.e., those with lower grades) benefit from SA quality and PA quality so that they use more calculation strategies (Ross et al., 1999; Sadler & Good, 2006).
H2a: Strategy variability has a positive effect on performance.
H2b: Particularly low-performing students (i.e., those with lower grades) benefit for their performance from using more calculation strategies.

Additionally, we explore the mediation effect / indirect path from PA quality and SA quality through calculation strategies on performance, for which there are no previous findings.

Methodology

In order to be able to record the use of PASA in real learning settings as validly as possible, lessons were video-recorded and analysed using a coding system and a highly inferential rating system.

Sample and Data Collection

For the study, we selected school classes at the fourth-year primary school level from German-speaking Switzerland. In six cantons (Aargau, Lucerne, Nidwalden, Obwalden, Uri, and Zug), we contacted a total of 423 school principals from 302 school communities, with a request to support the project and pass on the application forms and informational material to the teachers of their fourth-year classes. The education authorities of the respective cantons were also informed about the project and asked for their support in recruiting teachers. In the end, 52 school classes with 52 teachers and 634 students participated in the study. Most participating school classes are from the Canton of Lucerne (83%, 43 teachers). From the remaining five cantons, 1-3 school classes participated.

In return for their participation in the project, teachers could choose between written feedback on the video-recorded two lessons or further training at the school in a free one-hour presentation on the research topic. Participation in the study was voluntary, and the sample entails the self-selection of teachers and/or their school principals.

Due to the relatively time-consuming survey methods, the sample size was kept at 50 classes, as in other related studies (Beck et al., 2008). This sample size corresponds to the recommendations for multilevel analytical evaluations (Hox, 1998). According to this rule of thumb, a minimum of 30 level-2 units is required to test the fixed effects of level-2 factors (here, teachers or teaching characteristics) on a dependent variable. With the recruitment of 52 teachers, we exceeded the target of 50 teachers. Among the students, the planned sample size of 1000 students could not be reached. The original sample consists of 922 students. In 24 school classes, we excluded students from levels other than fourth-year (717 students remaining). Fifty-four of the legal guardians refused to allow their child to participate.

Furthermore, we excluded those students from the sample who were ill during the data collection or were mainly taught by the special education teacher, leaving 634 students for data analysis. Due to missing data in the rating variables of PASA, there were 384 students included for the main analyses of this paper. Whereas this resulted in a smaller sample size, there was still sufficient power (.84) to detect a relatively small effect (r = .15).

Data collection was carried out in the first half of 2017. Two lessons in mathematics were video-recorded by each of the 52 participating teachers. In order to ensure comparable framework conditions, the teaching topic "Introduction to whole-number-based strategy use for division problems" was specified. The topic is particularly suitable for the implementation of PASA. The whole-number-based strategy use for division problems enables a variety of approaches through a large number of possible calculation strategies. According to Schulz and Leuders (2018), the calculation strategies range from cognitively demanding thought processes (breaking down the divisor or the dividend) to rudimentary strategies (e.g., successive addition and subtraction). The potential for different calculation strategies enables the learners to be aware of their ways of thinking (SA) and to share, discuss and determine each other's learning status (PA).

The whole-number-based strategy use on division problems is a learning content from the mathematics curriculum of the fourth-year primary class in (German) Switzerland. We assume that the video-recorded lessons represent the mathematics lessons from the same teachers in general, since subjective theories and teaching scripts can be regarded as relatively stable (Groeben et al., 1988; Pianta & Hamre, 2009). The teachers were not informed about the project topic until after the data collection. As in other video observation studies (Hugener et al., 2006; Krammer & Hugener, 2014), the video recordings were standardised according to a detailed camera script. The video recordings were made by trained research group members, with the camera always following the teacher. After the video recordings, the students answered, with the researcher, items in a questionnaire, while the teachers completed a different questionnaire in a separate room. Before the video recordings and after the topic, the students solved a performance test for division.

Both the teachers and the legal guardians of the pupils involved gave active approval for participation. The parents' written informed consent was obtained for all students participating in the study, and all data protection requirements were met. Students with no permission sat outside the range of the camera.

Participants

The overall study sample included 52 teachers from six Swiss cantons, who teach fourth-grade classes in mathematics and other subjects. The sample comprised 40 female teachers and 12 male teachers aged 36 years on average (SD = 10.6; range: 25–60 years), with an average work experience of 10.6 years (SD = 10.3; range: 1.5–39.0 years). They taught an average of 22.1 lessons of 45 minutes per week in their classes (SD = 5.1; range: 4–29 lessons of different subjects). With 76.9% female teachers in the study, the proportion of women at this school level corresponds to the Swiss average of 76% in 2018 (Federal Statistical Office, 2021).

The sample of 634 participating students consisted of 315 girls and 315 boys; there was no gender information for four of the children. 35.6% of the students spoke a language other than (Swiss) German with at least one parent. The average age was 10.5 years (SD = 0.49; range: 9–12.6 years).

Measures

Quantity and Quality of PASA: A coding and a rating tool were developed to determine the frequency and duration (coding) and quality (rating) of PASA in the video recordings. There was a need to develop new instruments, because those currently available for monitoring formative assessment in the classroom (Gotwals et al., 2015; Oswalt, 2013) are not designed to measure frequency and quality.

The development of the instruments was guided by both the theory and the data (Pauli, 2012). The detailed development of the coding and rating system, and the results, are reported in Buholzer et al. (2020). The coding was done in the form of time samples with intervals of 10 seconds. The codes were used to determine, among other things, the video sections on PASA in the two math lessons. Ten percent of the video recordings were coded from two project staff members. The coders achieved high values of reliability (Interrater, PA quantity: Cohen's kappa of 0.95; SA quantity: Cohen's kappa of 0.89).

The quality was then rated for all video sections for which the occurrence of PASA had been coded before. The quality of PASA was assessed based on three items:

(1) Level of challenge: The assessment task is cognitively activating and demanding. Students must describe and grade their (SA) or other students (PA) learning process. Students have to justify their assessments.
(2) Guidance, implementation, and teacher support: The teacher supports and guides the assessment of the students. The instructions for the assessment are clearly formulated.
(3) Use of the information gained for further learning and teaching: A discussion takes place on the assessment results. The results are used for further learning (decisions).

The items were rated on a four-level scale ranging from "not present" (0), "low" (1), "medium" (2), to "high" (3). For each item, the level of proficiency was precisely defined and recorded in a rating manual. The rating was performed with the qualitative data analysis program MAXQDA (version 2018). Ten percent of the video material was rated by two people (Interrater, PA quality: ICC of 0.93; SA quality: ICC of 0.87). For double-rated values, the average value is taken.

Low-Performing and High-Performing Students: To determine who are the low-performing and high-performing students, we used the mathematics grade of the last semester rounded to half numbers. According to the grading system in Switzerland, high numbers mean good grades.

Performance Test: The performance test for the division was developed in cooperation with math didactics experts and based on specialist didactic literature (e.g., Schulz, 2015). The test items include, among other things, seven division tasks with graded levels of difficulty. The performance test was piloted in two primary school classes (150 students). The students completed the performance test twice, that is, before and after the lessons on the topic. For a standardised test correction, an evaluation manual was written. The following values were assigned to the tasks: 0 = wrong, 0.5 = partially correct, 1 = correct. Ten percent of the performance tests were double coded by the project staff and trained assistants (Interrater for correction of tasks, ICC of 0.99). For the double-coded values, the code of the project staff member was taken for the data set.

Variability of Calculation Strategy as Metacognitive Process: For each task of the performance test, the used whole-number-based calculation strategy was recorded, regardless of whether it was solved correctly. A distinction was made between the following calculation strategies: 1.) Column-based decomposition of dividend, 2.) Multiplicative decomposition of dividend, 3.) Decomposition of the divisor, 4.) Multiplicative decomposition of the divisor, 5.) Analogical transfer (tenfold analogy, analogue task, change in the same direction), 6.) Repeated addition, 7.) Mental calculation, 8.) Written calculation strategy, 9.) Drawing, 10.) Mixed, 11.) Others, 12.) Zero=blank field (Schulz, 2015). For each calculation strategy, examples of the individual strategy and coding rules were recorded in a manual. Ten percent of the performance tests were double coded (Interrater for strategy coding, Cohen's kappa of 0.91). Since the performance test was carried out twice, values of the strategies used are available for two measurement points.

Analysing of Data

Given the grouped structure of the data due to clustering in school classes, multilevel models were used for all research questions. More precisely, in all regression analyses, a random intercept for school class was included. The diagram in Figure 1 describes the moderated mediation of interest.

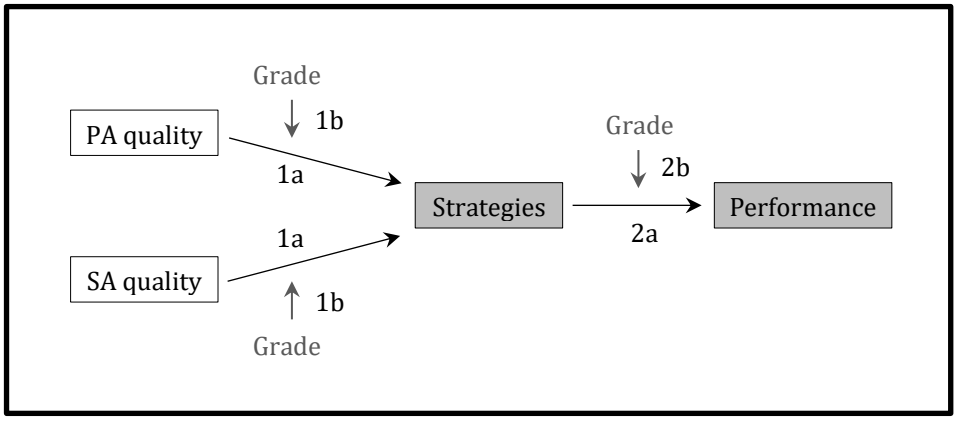

Figure 1. Path diagram of the moderated mediation.

We assume that PA quality and SA quality affect the metacognitive processes, operationalised as the variability of calculation strategies (direct paths, Hypothesis H1a), which affects performance (direct path, Hypothesis H2a). Both effects are moderated by the achievement level of the students, measured using the grade, meaning that the effects change depending on grade. Specifically, we assume that the effect is larger for students with lower grades (see Hypothesis H1b and H2b).

We estimated two models - Model 1 describes the direct path 1a and the moderation effect of grade 1b, and Model 2 describes the direct path 2a and the moderation effect of grade 2b.

The first model is used to estimate the moderated effects of the level 2 predictors SA quality and PA quality on the level 1 mediator use of more calculation strategies. The predictor variables in the model are: SA quality, PA quality, grade, the product of SA quality and grade, the product of PA quality and grade, and strategies T1. Strategies T1 is the measurement of the variability of strategies at time point 1, which is used to control the baseline level of the variability of strategies. Gender was included as a covariate. We are interested in the variability of strategies over time caused by high SA quality and PA quality.

The second model is used to estimate the moderated effect of the level 1 mediator strategies on the dependent variable of performance, controlled for the effect of SA quality and PA quality. Performance was measured by student performance in the division test at time point 2. Performance in the division test at time point 1 was included as a predictor, since we are interested in the performance gain. Thus, the predictor variables in Model 2 are: strategies (T2), SA quality, PA quality, grade, gender, the product of SA quality resp. PA quality and grade, the product of strategies and grade and performance (T1).

All analyses were conducted using R (R Core Team, 2020), and for the multilevel analyses, the packages "lme4" (Bates et al., 2015) and "lmerTest" (Kuznetsova et al., 2017) were used. Before running the multilevel analyses, all variables were scaled to have zero mean and unit variance. For conducting a multilevel regression, assumptions of normality and homogeneity of variances were checked. Normality was investigated by inspecting the histogram of residuals. One-way ANOVA and ICC(1) were calculated to test for non-independence of both strategies and performance.

Results

Descriptive Analysis

Quantity and Quality of PASA: PA was implemented by 88%, i.e., 46 of the 52 teachers. On average, 9.26 minutes (SD = 7.19; range: 0.83-33.17 minutes) was spent on PA during the two lessons. These 9.26 minutes make up 11.37% (SD = 8.89; range: 0.95-39.48%) of the mathematics teaching time. The average mathematics teaching time is 81.58 minutes (SD = 6.41, range: 60.66-92.00 minutes).

SA was observed for 40 teachers (77%) during an average of 1.10 minutes (SD = 0.88; range: 0.17-3.33 minutes) or 1.36% of the mathematics teaching time. The SA video clips were too short (less than 10 seconds) for two teachers, so that no quality rating could be made.

The rating results show that with M = 1.70, there is average PA quality, and a low SA quality of M = 1.14 (see Table 1).

Regarding the items, the teachers support and implement PASA relatively well. The cognitive demand of PASA is, on average, rather low, especially for SA. Furthermore the information gained from the assessments is rarely used for further learning.

Table 1. Items of PASA Quality.

No	Item	PA		SA	
		M	**SD**	**M**	**SD**
1.	Level of Challenge	1.73	0.77	1.04	.59
2.	Guidance, Implementation and Teacher Support	1.98	0.80	1.30	.65
3.	Use of Information Gained for Further Learning and Teaching	1.39	0.90	1.10	.98
	Average Quality	**1.70**	**0.74**	**1.14**	**0.53**

Note: SA n = 38, PA n = 46. Rating Scale: 0 = not present, 1 = low, 2 = medium, 3 = high.

Variability of Calculation Strategies: 617 of the 634 participating students took the performance test with the seven division tasks at the second measurement point. Students most frequently used the calculation strategy of column-based decomposition of dividend for 56.7% of the division tasks. This strategy was also most often taught in the two introductory lessons. The students used mental calculations for 2.1% and written calculation strategies for 1.2% of the tasks. The mixed, analogical transfer methods, multiplicative decomposition of the dividend, decomposition of dividend, and repeated addition were used for less than 1% of the tasks. 2.9% of the tasks were solved with strategies that could not be further determined. The students never applied the strategies of multiplicative decomposition of divisor and drawings. 32.3% of the task fields remained empty. The number of calculation strategies used in the performance test was calculated for each student. On average, students used 1.32 strategies at time point 2 (see Table 2).

Performance: Since performance is measured as the sum of seven binary variables, Table 2 shows that students on average solved half of the test items correctly (M = 3.50).

Table 2. Sample sizes, correlations, omega, means and standard deviations of predictors, mediator and outcome variable.

Variable	n	1	2	3	4	5	6	7	M	SD
Predictors (Level 2)										
1 PA Quality	46	-							1.70	0.74
2 SA Quality	38	.25**	-						1.14	0.53
Mediator (Level 1)										
3 Strategies (T2)	634	.09*	.05	-					1.32	0.63
Moderator (Level 1)										
4 Grade	616	.02	.04	-.02	-				5.00	0.64
Outcome (Level 1)										
5 Performance (T2)	617	.00	.00	.00	.53**	-			3.50	2.12
Covariates (Level 1)										
6 Strategies (T1)	634	.10*	.02	.11**	26**	.39**	-		1.10	0.80
7 Performance (T1)	622	.08	-.01	.07	.37**	.51**	.41**	-	1.03	1.41
8 Gender	630	-.07	-.05	-.02	.00	.09*	.07	.00	0.50	-

Note: * p<.05, two-tailed, ** p<.01, two-tailed. Gender is coded with 0 = male and 1 = female. Note that all correlations reported here were calculated at the individual student level (Level 1).
Internal consistency of ratings: Omega (SA Quality) = .84, Omega (PA Quality) = .92

Effects of PASA Quality on Student Variables

The ICC(1) for strategies is 0.00, with the ANOVA not being significant ($F(1,615) = 0.20$, $p = 0.65$) and for performance, the ICC(1) is 0.02 with the ANOVA being significant ($F(1,615) = 6.07$, $p = 0.01$). For consistency, we decided to run multilevel analyses for both regression models, despite the mediator strategies being independent of class.

Both models were tested against their corresponding null model – a random intercept model without any predictors. Both chi-square tests yielded a significant result (χ 21(7), = 23.08, (χ 22(9) = 209.48).

The results of both multilevel model estimations can be obtained from Table 3.

Table 3. Regression coefficients and standard errors for multilevel regression analysis.

	Model 1 (Strategies T2 as DV)		Model 2 (Performance T2 as DV)	
	β	***SE***	***β***	***SE***
PA Quality	.08	.07	.03	.10
SA Quality	-.01	.07	-.05	.08
PA Quality: Grade	-.05	.04	-.02	.03
SA Quality: Grade	-.12*	.04	.00	.03
Grade	.00	.05	.38**	.04
Strategies (T1)	.08	.05	-	-
Strategies (T2)	-	-	-.01	.03
Strategies (T2): Grade	-	-	-.10**	.03
Performance (T1)	-	-	.27**	.03
Gender	-.12*	.04	.07 *	.03

Note: n= 384, * $p < .05$, two-tailed, ** $p < .01$, two-tailed.
Gender is coded with 0 = male and 1 = female.

The left column shows the results for the regression on the mediator strategies T2 (path 1a). The data did not support our Hypothesis 1a, since the two direct-path main effects, PA quality and SA quality on strategies, were not evident. Across all students, there is no effect of PA quality or SA quality on strategies. However, the interaction of SA quality with grade (Hypothesis/path 1b) was significant. The interaction is negative, meaning that the lower the grade, the higher the influence of SA quality on strategies. These findings align with our Hypothesis 1b, that low-performing students benefit from high SA quality by using more strategies.

The right column of Table 3 shows the results for the regression on the dependent variable performance (path 2a). Again, there is no direct-path main effect of 2a strategies on performance, which means our Hypothesis 2a could not be supported. Across all students, the number of strategies used does not have any particular influence on student performance. Nevertheless, the interaction of strategies with grade was found to be significant. This interaction is consistent with Hypothesis 2b, that low-performing students benefit from using more strategies.

Regarding the covariates grade, it is not surprising that grade has a very substantial influence on performance. The observed influence of gender on the two dependent variables is too weak to be interpreted without a prior hypothesis.

The moderated mediation results are summarised in Figure 2, where the coefficients are the results for the interaction terms (moderation).

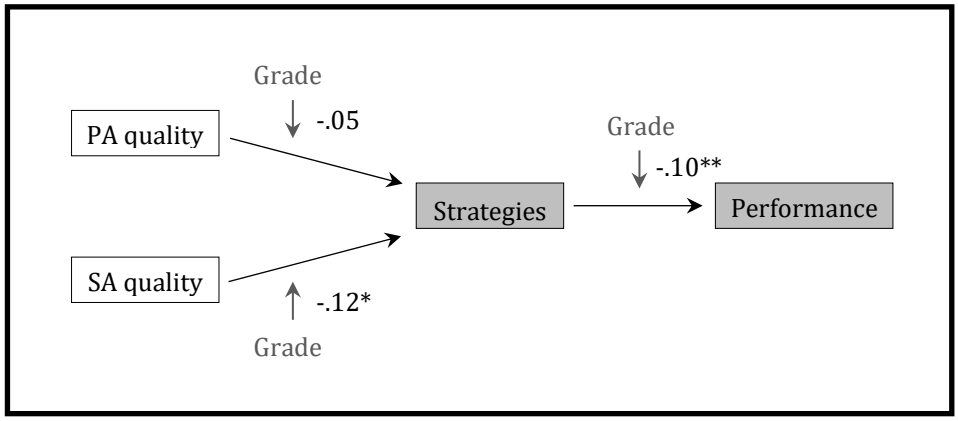

Figure 2. Path diagram of the moderated mediation with results for the interaction terms.

Subsequent analyses were only run for the model with SA quality as a predictor since here, the interaction term was significant on a .05 level. To further investigate the nature of this moderated mediation, we ran two additional analyses. First, we tested the overall mediation effect for four different values of grade (3, 4, 5, and 6), using the R package "mediate" (Tingley et al., 2014). The results only show a significant average causal mediation effect (ACME) for low-performance students with grade = 3 (ACME = 0.11, p = 0.04). The other three grade levels yield non-significant ACME's (see Appendix 1).

We further examined the moderation effect of the two single paths by conducting simple-slope analyses for the two significant moderation effects from Table 3. Figures 3 and 4 show four lines for the slopes of grades 3, 4, 5, and 6. Figure 3 clearly shows that only for low-performing students with grade 3 (dark line), do higher SA quality values increase strategies. The effect is almost zero for the other groups, being even slightly negative for high-performing students (grade = 6).

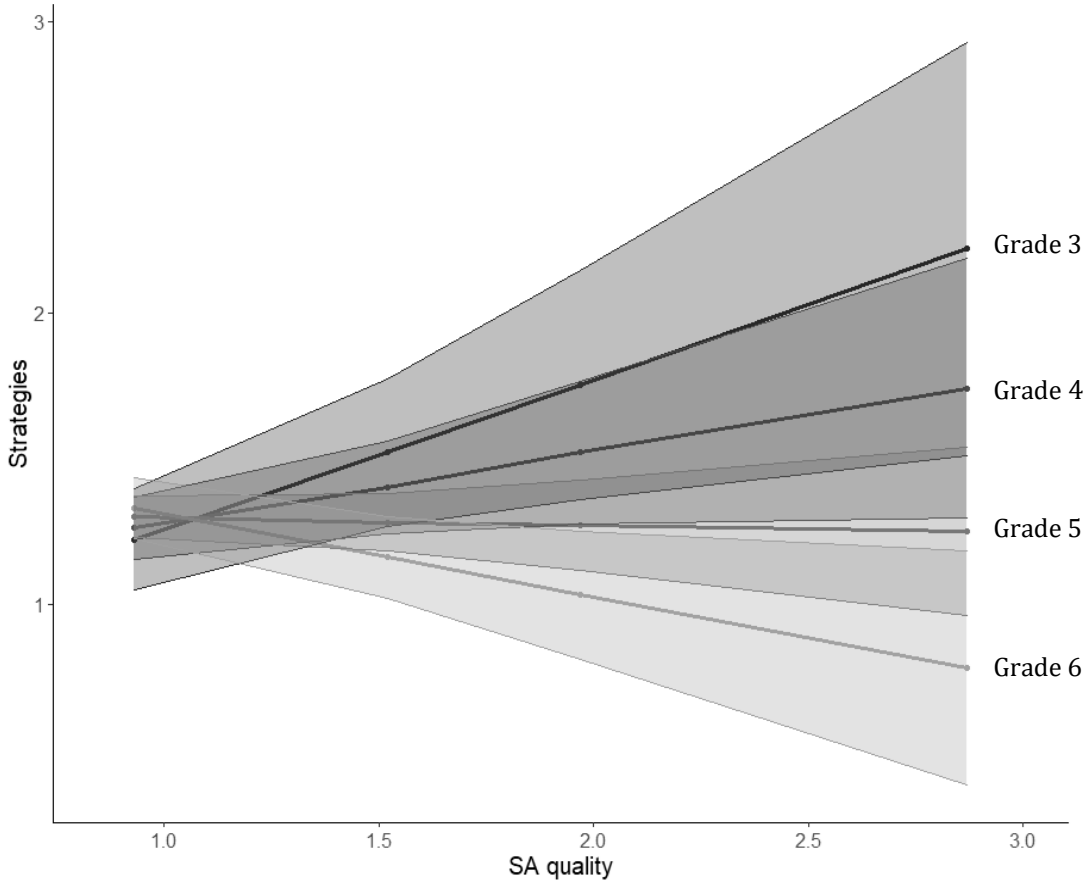

Figure 3. Simple slopes of the effect of SA quality on strategies for different grades.

The coefficients for the simple slopes are: 3: .37, 4: .18., 5: -.02, 6: -.20*.*
*p < .10, two-tailed, * p < .05, two-tailed*

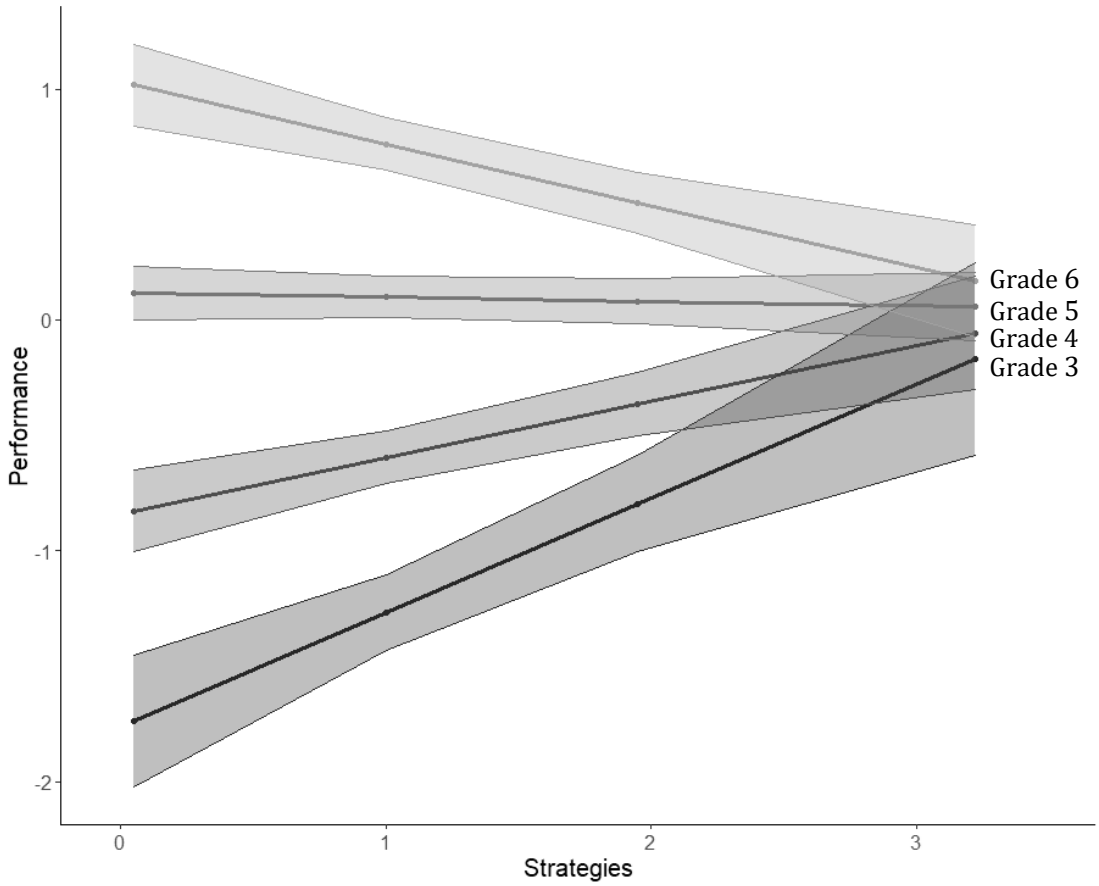

Figure 4. Simple slopes of the effect of strategies on performance for different grades.

The coefficients for the simple slopes are: 3: .31, 4: -.15*, 5: -.02, 6: -.17*,p < .10, two-tailed, * p < .05, two-tailed*

Figure 4 shows how different low- and high-performing students benefit from using more strategies, which only helps low-performing students to achieve better results in the performance test (dark line). Again, for medium-performing students, strategies barely affect performance, and for high-performing students, it even has a negative effect.

Discussion

This study aimed to determine how often and with what quality PASA occurs in everyday mathematics instruction, which students benefit, and what role calculation strategies play. Unlike many other studies on PASA, this study records real in-class behaviour, external observers rate the quality of PASA and relate it to student performance.

In the present study, the percentage of teachers implementing PASA in the classroom is relatively high (PA: 88%, SA: 77%), compared to the few studies, which all (in any event) report teacher self-evaluations (Cheng & Wang, 2007; Schmidt, 2020). The average duration of PASA (11.37% or 1.36% of the mathematics teaching time) is short, but reveals a large range of variation. Comparative values from other studies are not known.

The results of the quality rating indicate a low to medium quality of PASA. On the four-point scale (0 - 3), the quality of PA was rated higher (M = 1.70) than SA (M = 1.14). Of the three quality aspects investigated (level of challenge, teacher support, and use for further learning), the level of challenge, i.e., the cognitive activation of PASA, is, on average, rather low.

These results are largely consistent with the findings of the observational ratings of Gotwals et al. (2015) and Oswalt (2013), who also found quality scores in the lower scale range for PASA. By contrast, findings based on teacher self-reports (e.g., Altmann et al., 2010; Schmidt, 2020) yield much more positive results.

A direct path from the quality of PASA to the use of different calculation strategies and further, to the mathematics performance, could not be found. Unlike other studies (e.g. Bot, 2020; Ohadugha et al., 2020), including the meta-studies of Brown and Harris (2013), Double et al. (2019), Graham et al. (2015), and Sanchez et al. (2017), which found effects of PASA on performance, the present study was unable to show effects for all learners. In addition to the studies that showed effects, the meta-study from Brown and Harris (2013) contained studies with no or small effects, where the assumption

is that the quality is too low to show an overall effect on performance. The present study, with the missing main effects for all students, confirms this assumption.

The multilevel analysis results provide initial evidence that SA (rated within everyday mathematic teaching) is beneficial to some students. The interaction effects suggest that low-performing students (i.e. those with lower grades) benefit from quality SA in expanding their repertoire of strategies. One explanation for the different effects on the students could be that SA does not take on the same meaning for all learners. SA targeted all learner encounters with heterogeneous learning prerequisites in terms of mathematical knowledge and strategy use and can therefore not trigger metacognitive processes equally for all learners (Panadero et al., 2016). According to Siegler (2007), the development of strategies occurs over different phases and not at the same rate for all students. While some learners are introduced to the strategies, others are already familiar with different strategies, have tested their effectiveness, and adopted the appropriate strategy. Due to these heterogeneous conditions, differential effects of PASA are to be expected. The present study shows that expanding their repertoire of strategies helps low-performing students improve their performance. SA triggers metacognitive thinking processes in low-performing students. Expanding the repertoire of strategies leads to an adaptive adjustment of one's abilities and tasks, thus indirectly improving mathematical performance. The results confirm the research findings of Ross et al. (1999) and Sadler and Good (2006), who also demonstrated the benefits of SA, primarily for lower performers. Low-performing students often have lower metacognitive knowledge (Miller & Geraci, 2011), stimulated and enhanced by high quality SA. Although the found interaction effects were relatively small, they can matter in the long run. That is, single events accumulate over time into important effects, even when the measured effects at a certain time are small (Funder & Ozer, 2019).

The same effects could not be proven for PA. The quality of PA was rated higher than for SA, but there was no effect on the low-performing students. The learners in PA deal with their peers' solution paths and results (Strijbos & Wichmann, 2017). In contrast, SA can be viewed as learners giving feedback to themselves, which eliminates corrective and sometimes hurtful feedback from peers and teachers. Furthermore, PA can be more challenging for low-performing students than SA; PA requires knowledge and skills of the core task and the implementation of PA. Low-performing students rarely have the appropriate mathematical knowledge to place themselves in a peer's thinking process and understand the solution path (Strijbos & Wichmann, 2017). The impact of formative PA depends not only on the quality and content of the feedback received, but also on the recipient's feedback processing depth. The processing of PA is influenced by cognitive, social, and motivational factors, including the perceived expertise of the sender or beliefs about PA (Deiglmayr, 2018). Since this information on the individual interactions and the people involved is missing, further explanations for the missing effects of PA are difficult to provide.

High-performing students in this study do not benefit from either quality PA or SA, which does not necessarily mean that they do not benefit from PASA quality for their learning process. At the end of the learning topic, high-performing students had already completed the phase of trying out different strategies. They consolidated one strategy for themselves at the second performance measurement.

Conclusion

The occurrence, quality, and effects of PASA were investigated in everyday lessons by coding video-recorded lessons and assessing them with external raters. The results provide evidence of real classroom practice and are thus ecologically more valid than the (quasi-)experimental studies usually conducted in this field of research.

PASA is carried out in everyday teaching by a comparatively large number of teachers, but only for a relatively short duration and tends to be of medium (PA) or low (SA) quality. However, the differences between teachers concerning time and quality are large.

Effects of SA on low-performing students are evident in the case of qualitatively well-conducted SA. Good quality SA has a positive effect on low-performing students' use of different calculation strategies. Metacognitive thinking, operationalised as the use of different calculation strategies, positively affects low-performing students' performance.

From the results, it can be concluded that SA, if implemented qualitatively well, positively affects the learning process. The overall effectiveness of SA, and possibly PA, could be increased by more frequent occurrence and good quality implementation. The high variance of the results of the quantity and quality of PASA confirm the conclusions of Grob et al. (2019) that more formative assessment practices are possible and desirable.

With the use of good quality SA, low-performing students can be supported. So, good quality SA may help to reduce the gap between low- and high-performing students.

Recommendations

Regarding PASA rating, other studies show that the observation of cognitive activation, in our study contained in the level of challenge, requires more observation time. Due to the wider variety of cognitive activation and further possibilities of influence, Praetorius et al. (2014) recommend nine hours of observation time. Furthermore, new research findings suggest the need for a subject-specific concretisation of the cognitive activation, especially regarding the quality,

completeness, and type of content (Reusser & Pauli, 2021). A more subject-didactic orientation of the PASA observation instrument with a longer observation period should be considered in further studies.

According to the overlapping waves theory (Chen & Siegler, 2000; Siegler, 2002), different strategies are always available for selection throughout development. Thus, it can be assumed or should be examined in further studies, whether good SA also reveals the same effects at other age levels and with other subjects, where adaptive strategy selection can effectively improve performance. Furthermore, it would be useful to examine whether good SA reveals an advantage in high or middle-performing students, for example, by measuring performance and strategy variability at different points in the learning process. It could well be the case that high- or middle-performing students benefit from good quality SA more at the beginning of the learning process and not at the end, where we measured the variability of strategy use and the performance.

Concerning teaching practice, it could well be the case that for PASA to have a strong impact, it needs to be an active part of the curriculum with an overall high-quality level. The overall effectiveness of PASA could thus be increased by the more frequent occurrence of good quality PASA. As Heritage (2020) mentioned, the implementation of PASA is an issue of professional learning. Teachers need to be able to design lesson structures and routines so that opportunities for PASA can be built into the rhythm of teaching and learning.

Limitations

The project sample can be considered relatively large compared to other video studies (Gotwals et al., 2015; Lyon et al., 2019; Oswalt, 2013; Ruiz-Primo & Furtak, 2006). However, the size is still too limited for in-depth analyses of specific emerging patterns of formative assessment. In addition, positive selection can be assumed in the sample's composition, since the participating teachers had to agree to be videorecorded while teaching.

Concerning the study design, previous studies that demonstrated the effects of PASA on performance were (quasi-)experimental, resulting in large differences between groups, and performance measures adapted to the intervention. Within our study, differences in PASA between classes were smaller, and consequently, many other aspects might have also influenced performance. Furthermore, in contrast to (quasi-)experimental studies, the PASA practices in everyday mathematics teaching are not linked to the outcomes under investigation. Therefore, large effects cannot be expected in this context (Kraft, 2020).

Causal statements are also problematic, because the present study is not based on an experimental design. Therefore, statements can only be made about the correlations between quality SA and an expansion of the repertoire of strategies among low-performing students.

The reported results on PASA quality in the classroom apply only to fourth-year mathematics classes in (central) Switzerland. Generalizability to other subjects and school levels is very limited, as there is evidence that formative assessment is quite domain-specific (Maier, 2011) and may also vary across school levels (Bürgermeister, 2014).

Acknowledgements

The authors would like to thank Prof. Dr Matthias Baer, Hanni Lötscher, and Andrea Häfliger from the University of Teacher Education Lucerne for their support in planning and conducting the study. This article has benefited from Dr Brian Bloch, Dr Stella Bollmann, Dr Simon Breil and Dr Vu Thi Thao.

Funding

This work was supported by the Swiss National Science Foundation SNF [SNF-Project Nr. 100019 169771].

Authorship Contribution Statement

Zulliger: conceptualisation, design, data analysis/interpretation, drafting the manuscript, writing, statistical analysis. Buholzer: conceptualisation, design, data analysis/interpretation, drafting the manuscript, writing, data acquisition, securing funding, supervision, final approval. Ruelmann: data acquisition, critical revision of the manuscript.

References

Alqassab, M. (2016). *Peer feedback provision and mathematical proofs: Role of domain knowledge, beliefs, perceptions, epistemic emotions, and peer feedback content* [Doctoral thesis, Ludwig-Maximilians University]. Ludwig-Maximilians University. https://bit.ly/3zaTyMK

Altmann, P. C., Fleming, P. B., & Heyburn, S. L. (2010). *Understanding and using formative assessments: A mixed methods study of assessment for learning adoption.* Vanderbilt University. https://bit.ly/3HocNVZ

Andrade, H. L. (2010). Students as the definitive source of formative assessment: Academic self-assessment and the self-regulation of learning. In H. L. Andrade & G. J. Cizek (Eds.), *Handbook of formative assessment* (pp. 90–105). Routledge. https://doi.org/10.4324/9780203874851

Andrade, H. L. (2019). A critical review of research on student self-assessment. *Frontiers in Education,* (4), Article 87. https://doi.org/10.3389/feduc.2019.00087

Andrade, H. L., & Valtcheva, A. (2009). Promoting learning and achievement through self-assessment. *Theory into Practice*, *48*(1), 12–19. https://doi.org/10.1080/00405840802577544

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1-48. https://doi.org/10.18637/jss.v067.i01

Beck, E., Baer, M., Guldimann, T., Bischoff, S., Brühwiler, C., Müller, P., Niedermann, R., Rogalla, M., & Vogt, F. (2008). *Adaptive Lehrkompetenz: Analyse und Struktur, Veränderbarkeit und Wirkung handlungssteuernden Lehrerwissens* [Adaptive teaching competence: analysis and structure, changeability and effect of action-controlling teacher knowledge]. Waxmann.

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2004). Working inside the black box: Assessment for learning in the classroom. *Phi Delta Kappan*, *86*(1), 8–21. https://doi.org/10.1177/003172170408600105

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

Bot, T. D. (2020). On categories of mathematics teachers' classroom characteristics and perceived influence on effective mathematics teaching in secondary schools in Plateau state, Nigeria. *European Journal of Mathematics and Science Education, 1*(2), 121-130. https://doi.org/10.12973/ejmse.1.2.121

Boud, D., Lawson, R., & Thompson, D. G. (2013). Does student engagement in self-assessment calibrate their judgement over time? *Assessment & Evaluation in Higher Education*, *38*(8), 941–956. https://doi.org/10.1080/02602938.2013.769198

Brookhart, S. M., Moss, C. M., & Long, B. A. (2010). Teacher inquiry into formative assessment practices in remedial reading classrooms. *Assessment in Education: Principles, Policy & Practice*, *17*(1), 41–58. https://doi.org/10.1080/09695940903565545

Brown, G. L., & Harris, L. R. (2013). Student self-assessment. In J. H. McMillan (Ed.), *SAGE handbook of research on classroom assessment* (pp. 367–393). SAGE. https://doi.org/10.4135/9781452218649.n21

Buholzer, A., Baer, M., Zulliger, S., Torchetti, L., Ruelmann, M., Häfliger, A., & Lötscher, H. (2020). Formatives Assessment im alltäglichen Mathematikunterricht von Primarlehrpersonen: Häufigkeit, Dauer und Qualität [Formative assessment in the everyday mathematics teaching of primary teachers: Frequency, duration and quality]. *Unterrichtswissenschaft*, *48*(4), 629–661. https://doi.org/10.1007/s42010-020-00083-7

Bürgermeister, A. (2014). *Leistungsbeurteilung im Mathematikunterricht: Bedingungen und Effekte von Beurteilungspraxis und Beurteilungsgenauigkeit* [Performance assessment in mathematics education: conditions and effects of assessment practice and assessment accuracy]. Waxmann.

Cardelle-Elawar, M. (1995). Effects of metacognitive instruction on low achievers in mathematics problems. *Teaching and Teacher Education*, *11*(1), 81–95. https://doi.org/10.1016/0742-051X(94)00019-3

Chen, Z., & Siegler, R. S. (2000). II. Overlapping waves theory. *Monographs of the Society for Research in Child Development*, *65*(2), 7–11. https://doi.org/10.1111/1540-5834.00075

Cheng, L., & Wang, X. (2007). Grading, feedback, and reporting in ESL/EFL classrooms. *Language Assessment Quarterly*, *4*(1), 85–107. https://doi.org/10.1080/15434300701348409

Cizek, G. J. (2010). An introduction to formative assessment: History, characteristics, and challenges. In H. L. Andrade & G. J. Cizek (Eds.), *Handbook of formative assessment* (pp. 3–17). Routledge. https://doi.org/10.4324/9780203874851

Decristan, J., Klieme, E., Kunter, M., Hochweber, J., Büttner, G., Fauth, B., Hondrich, A. L., Rieser, S., Hertel, S., & Hardy, I. (2015). Embedded formative assessment and classroom process quality. *American Educational Research Journal*, *52*(6), 1133–1159. https://doi.org/10.3102/0002831215596412

Deiglmayr, A. (2018). Instructional scaffolds for learning from formative peer assessment: Effects of core task, peer feedback, and dialogue. *European Journal of Psychology of Education*, *33*(1), 185–198. https://doi.org/10.1007/s10212-017-0355-8

Double, K. S., McGrane, J. A., & Hopfenbeck, T. N. (2019). The impact of peer assessment on academic performance: A meta-analysis of control group studies. *Educational Psychology Review*, *32*(2), 481–509. https://doi.org/10.1007/s10648-019-09510-3

Fagginger Auer, M. F., Hickendorff, M., & van Putten, C. M. (2016). Solution strategies and adaptivity in multidigit division in a choice/no-choice experiment: Student and instructional factors. *Learning and Instruction*, *41*, 52–59. https://doi.org/10.1016/j.learninstruc.2015.09.008

Fauth, B., Decristan, J., Rieser, S., Klieme, E., & Büttner, G. (2014). Grundschulunterricht aus Schüler-, Lehrer- und Beobachterperspektive: Zusammenhänge und Vorhersage von Lernerfolg [Primary school teaching from pupil, teacher and observer perspectives: Interrelationships and prediction of learning success]. *Zeitschrift für Pädagogische Psychologie*, *28*(3), 127–137. https://doi.org/10.1024/1010-0652/a000129

Federal Statistical Office. (2021). *Lehrkräfte nach Bildungsstufe (öffentliche Schulen)* [Teachers by education level (public schools)]. Federal Statistical Office. https://bit.ly/3sUpTq1

Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science*, *2*(2), 156–168. https://doi.org/10.1177/2515245919847202

Geary, D. C., & Hoard, M. K. (2005). Learning disabilities in arithmetic and mathematics. In J. I. D. Campbell (Ed.), *Handbook of mathematical cognition* (pp. 253–267). Psychology Press.

Gersten, R., Chard, D. J., Jayanthi, M., Baker, S. K., Morphy, P., & Flojo, J. (2009). Mathematics instruction for students with learning disabilities: A meta-analysis of instructional components. *Review of Educational Research*, *79*(3), 1202–1242. https://doi.org/10.3102/0034654309334431

Geurten, M., & Lemaire, P. (2019). Metacognition for strategy selection during arithmetic problem-solving in young and older adults. *Neuropsychology, Development, and Cognition. Section B, Aging, Neuropsychology and Cognition*, *26*(3), 424–446. https://doi.org/10.1080/13825585.2018.1464114

Gielen, S., Peeters, E., Dochy, F., Onghena, P., & Struyven, K. (2010). Improving the effectiveness of peer feedback for learning. *Learning and Instruction*, *20*(4), 304–315. https://doi.org/10.1016/j.learninstruc.2009.08.007

Gotwals, A. W., Philhower, J., Cisterna, D., & Bennett, S. (2015). Using video to examine formative assessment practices as measures of expertise for mathematics and science teachers. *International Journal of Science and Mathematics Education*, *13*(2), 405–423. https://doi.org/10.1007/s10763-015-9623-8

Graham, S., Hebert, M., & Harris, K. R. (2015). Formative assessment and writing. *The Elementary School Journal*, *115*(4), 523–547. https://doi.org/10.1086/681947

Grob, R., Holmeier, M., & Labudde, P. (2019). Analysing formal formative assessment activities in the context of inquiry at primary and upper secondary school in Switzerland. *International Journal of Science Education*, *43*(3), 407-427. https://doi.org/10.1080/09500693.2019.1663453

Groeben, N., Wahl, D., Schlee, J., & Scheele, B. (1988). *Das Forschungsprogramm subjektive Theorien: eine Einführung in die Psychologie des reflexiven Subjekts* [The research programme subjective theories: an introduction to the psychology of the reflexive subject]. Francke.

Harris, L. R., & Brown, G. T. (2013). Opportunities and obstacles to consider when using peer- and self-assessment to improve student learning: Case studies into teachers' implementation. *Teaching and Teacher Education*, *36*, 101–111. https://doi.org/10.1016/j.tate.2013.07.008

Hartnett, J. (2007). Categorisation of mental computation strategies to support teaching and to encourage classroom dialogue. In J. Watson & K. Beswick (Eds.), *Mathematics: Essential research, essential practice* (pp. 345–352). MERGA.

Heinze, A., Star, J. R., & Verschaffel, L. (2009). Flexible and adaptive use of strategies and representations in mathematics education. *ZDM*, *41*(5), 535–540. https://doi.org/10.1007/s11858-009-0214-4

Heritage, M. (2020). Getting the emphasis right: Formative assessment through professional learning. *Educational Assessment*, *25*(4), 355–358. https://doi.org/10.1080/10627197.2020.1766959

Hickendorff, M., Torbeyns, J., & Verschaffel, L. (2019). Multi-digit addition, subtraction, multiplication, and division strategies. In A. Fritz, V. G. Haase, & P. Räsänen (Eds.), *International handbook of mathematical learning difficulties* (pp. 543–560). Springer. https://doi.org/10.1007/978-3-319-97148-3_32

Hill, T. (2016). Do accounting students believe in self-assessment? *Accounting Education*, *25*(4), 291–305. https://doi.org/10.1080/09639284.2016.1191271

Hox, J. J. (1998). Multilevel modeling: When and why. In I. Balderjahn, R. Mathar & M. Schader (Eds.), *Classification, data analysis, and data highways* (pp. 147–154). Springer. https://doi.org/10.1007/978-3-642-72087-1_17

Hugener, I., Pauli, C., & Reusser, K. (2006). *Dokumentation der Erhebungs- und Auswertungsinstrumente zur schweizerisch-deutschen Videostudie "Unterrichtsqualität, Lernverhalten und mathematisches Verständnis": 3. Videoanalysen* [Documentation of the survey and evaluation instruments for the Swiss-German video study "Teaching quality, learning behaviour and mathematical understanding": 3. video analyses.]. Gesellschaft z. Förd. Päd. Forsch. https://bit.ly/32QO1if

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

Kollar, I., & Fischer, F. (2010). Peer assessment as collaborative learning: A cognitive perspective. *Learning and Instruction*, *20*(4), 344–348. https://doi.org/10.1016/j.learninstruc.2009.08.005

Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher*, *49*(4), 241–253. https://doi.org/10.3102/0013189X20912798

Krammer, K., & Hugener, I. (2014). Förderung der Analysekompetenz angehender Lehrpersonen anhand von eigenen und fremden Unterrichtsvideos [Promoting the analytical competence of prospective teachers by means of their own and other people's teaching videos]. *Journal für LehrerInnenbildung*, *14*(1), 25–32. https://bit.ly/3HJT1Vp

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1-26. https://doi.org/10.18637/jss.v082.i13

Lindberg, S., Linkersdörfer, J., Lehmann, M., Hasselhorn, M., & Lonnemann, J. (2013). Individual differences in children's early strategy behavior in arithmetic tasks. *Journal of Educational and Developmental Psychology*, *3*(1), 192-200. https://doi.org/10.5539/jedp.v3n1p192

Lyon, C. J., Nabors Oláh, L., & Wylie, E. C. (2019). Working toward integrated practice: Understanding the interaction among formative assessment strategies. *The Journal of Educational Research*, *112*(3), 301–314. https://doi.org/10.1080/00220671.2018.1514359

Maier, U. (2011). Formative Leistungsdiagnostik in der Sekundarstufe I - Befunde einer quantitativen Lehrerbefragung zu Nutzung und Korrelaten verschiedener Typen formativer Diagnosemethoden in Gymnasien [Formative performance diagnostics in lower secondary schools - Findings from a quantitative teacher survey on the use and correlates of different types of formative diagnostic methods in grammar schools]. *Empirische Pädagogik*, *25*(1), 25–46.

Miller, T. M., & Geraci, L. (2011). Unskilled but aware: Reinterpreting overconfidence in low-performing students. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *37*(2), 502–506. https://doi.org/10.1037/a0021802

Ohadugha, R. O., Chukwuemeka, E. J., & Babatunde, A. E. (2020). Impact of peer-mediated learning on achievement and motivation in computer science among senior secondary school students in Minna Metropolis, Niger State. *Contemporary Educational Technology, 12*(1), ep263. https://doi.org/10.30935/cedtech/7629

Oswalt, S. G. (2013). *Identifying formative assessment in classroom instruction* [Doctoral thesis, Boise State University]. Boise State University Scholar Works. https://bit.ly/3qF256K

Panadero, E. (2016). Is it safe? Social, interpersonal, and human effects of peer assessment: A review and future directions. In G. T. L. Brown & L. R. Harris (Eds.), *Handbook of social and human conditions in assessment* (pp. 247–266). Routledge.

Panadero, E., Brown, G. L., & Strijbos, J. -W. (2016). The future of student self-assessment: A review of known unknowns and potential directions. *Educational Psychology Review*, *28*(4), 803–830. https://doi.org/10.1007/s10648-015-9350-2

Pantiwati, Y., & Husamah, H. (2017). Self and peer Assessments in active learning model to increase metacognitive awareness and cognitive abilities. *International Journal of Instruction*, *10*(4), 185–202. https://doi.org/10.12973/iji.2017.10411a

Pauli, C. (2012). Kodierende Beobachtung [Coding observation]. In H. de Boer & S. Reh (Eds.), *Beobachtung in der Schule - Beobachten lernen* [Observation at school - learn to observe] (pp. 45–63). Springer. https://doi.org/10.1007/978-3-531-18938-3_3

Pennequin, V., Sorel, O., Nanty, I., & Fontaine, R. (2010). Metacognition and low achievement in mathematics: The effect of training in the use of metacognitive skills to solve mathematical word problems. *Thinking & Reasoning*, *16*(3), 198–220. https://doi.org/10.1080/13546783.2010.509052

Pianta, R. C., & Hamre, B. K. (2009). Classroom processes and positive youth development: Conceptualizing, measuring, and improving the capacity of interactions between teachers and students. *New Directions for Youth Development*, *2009*(121), 33–46. https://doi.org/10.1002/yd.295

Ploegh, K., Tillema, H. H., & Segers, M. S. (2009). In search of quality criteria in peer assessment practices. *Studies in Educational Evaluation*, *35*(2-3), 102–109. https://doi.org/10.1016/j.stueduc.2009.05.001

Praetorius, A. -K. (2014). *Messung von Unterrichtsqualität durch Ratings* [Measuring teaching quality through ratings]. Waxmann.

Praetorius, A. -K., Pauli, C., Reusser, K., Rakoczy, K., & Klieme, E. (2014). One lesson is all you need? Stability of instructional quality across lessons. *Learning and Instruction*, *31*, 2–12. https://doi.org/10.1016/j.learninstruc.2013.12.002

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Reusser, K., & Pauli, C. (2021). Unterrichtsqualität ist immer generisch und fachspezifisch. Ein Kommentar aus kognitions- und lehr-lerntheoretischer Sicht [Teaching quality is always generic and subject-specific. A commentary from a cognitive and learning theory perspective.]. *Unterrichtswissenschaft*, *49*(2), 189–202. https://doi.org/10.1007/s42010-021-00117-8

Ross, J. A., Rolheiser, C., & Hogaboam-Gray, A. (1999). Effect of self-evaluation on narrative writing. *Assessing Writing*, *6*(1), 107–132. https://doi.org/10.1016/S1075-2935(99)00003-3

Ruiz-Primo, M. A., & Furtak, E. M. (2006). Informal formative assessment and scientific inquiry: Exploring teachers' practices and student learning. *Educational Assessment*, *11*(3-4), 237–263. https://doi.org/10.1080/10627197.2006.9652991

Sadler, P., & Good, E. (2006). The impact of self- and peer-grading on student learning. *Educational Assessment*, *11*(1), 1–31. https://doi.org/10.1207/s15326977ea1101_1

Sanchez, C. E., Atkinson, K. M., Koenka, A. C., Moshontz, H., & Cooper, H. (2017). Self-grading and peer-grading for formative and summative assessments in 3rd through 12th grade classrooms: A meta-analysis. *Journal of Educational Psychology*, *109*(8), 1049-1066. https://doi.org/10.1037/edu0000190

Schmidt, C. A. (2020). *Formatives Assessment in der Grundschule: Konzept, Einschätzungen der Lehrkräfte und Zusammenhänge* [Formative assessment in primary school: Concept, teachers' assessments and contexts] (1st ed.). Springer. https://doi.org/10.1007/978-3-658-26921-0

Schnell, R., Hill, P. B., & Esser, E. (2013). *Methoden der empirischen Sozialforschung* [Methods of empirical social research] (10th revised ed.). Oldenbourg Verlag.

Schulz, A. (2015). Wie lösen Viertklässler Rechenaufgaben zur Multiplikation und Division? [How do fourth graders solve multiplication and division problems?] In F. Caluori, H. Linneweber-Lammerskitten & C. Streit (Eds.), *Beiträge zum Mathematikunterricht 2015* (pp. 844–847). WTM. https://doi.org/10.17877/DE290R-16783

Schulz, A., & Leuders, T. (2018). Learning trajectories towards strategy proficiency in multi-digit division – A latent transition analysis of strategy and error profiles. *Learning and Individual Differences*, *66*, 54–69. https://doi.org/10.1016/j.lindif.2018.04.014

Schütze, B., Souvignier, E., & Hasselhorn, M. (2018). Stichwort – Formatives Assessment [Keyword - Formative assessment]. *Zeitschrift für Erziehungswissenschaft*, *21*(4), 697–715. https://doi.org/10.1007/s11618-018-0838-7

Siegler, R. S. (1996). A grand theory of development. *Monographs of the Society for Research in Child Development*, *61*(1-2), 266–275. https://doi.org/10.1111/j.1540-5834.1996.tb00550.x

Siegler, R. S. (2002). Microgenetic studies of self-explanation. In N. Granott & J. Parziale (Eds.), *Microdevelopment* (pp. 31–58). Cambridge University Press. https://doi.org/10.1017/CBO9780511489709.002

Siegler, R. S. (2007). Cognitive variability. *Developmental Science*, *10*(1), 104–109. https://doi.org/10.1111/j.1467-7687.2007.00571.x

Strijbos, J. -W., Narciss, S., & Dünnebier, K. (2010). Peer feedback content and sender's competence level in academic writing revision tasks: Are they critical for feedback perceptions and efficiency? *Learning and Instruction*, *20*(4), 291–303. https://doi.org/10.1016/j.learninstruc.2009.08.008

Strijbos, J. -W., & Wichmann, A. (2017). Promoting learning by leveraging the collaborative nature of formative peer assessment with instructional scaffolds. *European Journal of Psychology of Education*, *33*(1), 1–9. https://doi.org/10.1007/s10212-017-0353-x

Tingley, D., Yamamoto, T., Hirose, K., Keele, L., & Imai, K. (2014). Mediation: R package for causal mediation analysis. *Journal of Statistical Software*, *59*(5), 1-38. https://doi.org/10.18637/jss.v059.i05

Topping, K. J. (2009). Peer assessment. *Theory into Practice*, *48*(1), 20–27. https://doi.org/10.1080/00405840802577569

Waldis, M., Grob, U., Pauli, C., & Reusser, K. (2010). Der schweizerische Mathematikunterricht aus der Sicht von Schülerinnen und Schülern und in der Perspektive hochinferenter Beobachterurteile [Swiss mathematics teaching from the perspective of students and in the perspective of highly-inferental ratings]. In K. Reusser, C. Pauli & M. Waldis (Eds.), *Unterrichtsgestaltung und Unterrichtsqualität: Ergebnisse einer internationalen und schweizerischen Videostudie zum Mathematikunterricht* [Lesson design and teaching quality: Results of an international and Swiss video study on mathematics teaching] (pp. 171–208). Waxmann.

Wylie, C., & Lyon, C. (2013). *Using the formative assessment rubrics, reflection and observation tools to support professional reflection on practice*. Council of Chief State School Officers (CCSSO). https://bit.ly/3Jub5Em

Appendix

Confidence intervals and p-values of ACME's from moderated mediation analysis

	Grade = 3	**Grade = 4**	**Grade =5**	**Grade = 6**
CI	0.003 – 0.28	-0.005 – 0.08	-0.005 – 0.01	-0.0005 – 0.09
p	0.039	0.12	0.98	0.057