ZDM – Mathematics Education (2024) 56:815–830
https://doi.org/10.1007/s11858-024-01562-2

ORIGINAL PAPER

# On the added value of considering effects of generic and subject-specific instructional quality on students' achievements – an exploratory study on the example of implementing formative assessment in mathematics education

**Michael Besser**[1] · **Maike Hagena**[2] · **Thilo Kleickmann**[3]

Accepted: 6 March 2024 / Published online: 23 May 2024


**Abstract**
Formative Assessment (FA) is a promising teaching practice for supporting students' learning at school. However, implementing FA into teaching (mathematics) is challenging, effects of implementation of FA vary between empirical studies. Therefore, recent studies additionally consider instructional quality of teaching when analyzing effects of FA on students' learning. The current exploratory study builds on this idea and highlights the added value of considering both generic and subject-specific instructional quality (GIQ and SSIQ) when analyzing effects of FA: Based on a re-analyzation of data from the research project Co$^2$CA, 856 students from 39 mathematics classes of German middle schools took part in an intervention control-study aiming at implementing FA into mathematics education at school. On the student level, students' mathematics achievement is assessed by standardized tests before and after the intervention. During the intervention, on the class level GIQ and SSIQ are assessed by low- and high-inference ratings of 72 video-taped lessons. GIQ is based on the model of Three Basic Dimensions, SSIQ is based on the normative idea that mathematics teaching should offer opportunities for a deep understanding of subject matter and for building up competencies. Multi-level regression analyses highlight different (interaction) effects of GIQ and SIQ on students' performances. It becomes obvious: Considering GIQ and SIQ can have an added value for the better understanding of implementing FA into teaching (in mathematics education).

**Keywords** Formative assessment · Mathematics education · Instructional quality · Generic perspective · Subject-specific perspective · Students' achievement

## 1 Introduction

In his famous work, Hattie (2008) analyses the effects of "what actually works in schools to improve learning". He highlights that one of the most powerful impact factors of successful teaching is FA. Therefore, within recent years researchers investigated how to implement FA into teaching (of mathematics) best and how to support teachers in doing so successfully (Boström & Palm, 2023; Johnson et al., 2019; Schütze et al., 2017). A key result of these efforts is that understanding effects of FA at school implies not only analyzing the quality of implementation of this specific teaching practice itself but considering "global factors of high quality teaching and quality components of specific teaching practices" (Decristan et al., 2015, p. 1153) simultaneously. This is in line with theoretical and empirical work about the impact of learning processes in classroom (Good et al., 2009; Patrick et al., 2012; Seidel & Shavelson, 2007) as well as about the characteristics and influences of teachers' competences (Blömeke et al., 2015; Kunter et al., 2013) on students' development: It is the (generic and subject-specific) instructional quality being an important mediator between teachers' expertise and learning outcomes (Muijs et al., 2014; Roehring et al., 2012). Based on this, some very first studies have analyzed (interaction) effects of GIQ and FA on students' learning simultaneously (Decristan et

✉ Michael Besser
michael.besser@leuphana.de

[1] Leuphana University Lueneburg, Lueneburg, Germany

[2] Leibniz University Hannover, Hanover, Germany

[3] Kiel University, Kiel, Germany

al., 2015; Pinger et al., 2018), but studies considering both GIQ and SSIQ when investigating the quality of implementation of FA are not known to the authors. Starting from this desideratum the present exploratory study analyses the added value of considering (interaction) effects of GIQ and SSIQ on students' learning when implementing FA. Therefore, the paper is structured as follows: Within the theoretical part, central considerations about FA (Sect. 2) as well as about (FA in the context of) instructional quality (both GIQ and SSIQ) are outlined (Sect. 3). Within the empirical part, research question and hypotheses are given (Sect. 4) and methods (Sect. 5) as well as results (Sect. 6) are described. Finally, implications and limitations (Sect. 7) are discussed by pointing out the added value of considering GIQ and SSIQ in empirical educational research.

## 2 Formative assessment (in mathematics education)

Black and Wiliam (1998, p. 7) define FA as "all those activities undertaken by teachers, and/or by their students which provide information to be used as feedback to modify the teaching and learning activities" (see also Black & Wiliam, 2009). "All those activities" subsume different kinds of assessment (like teacher-directed assessment, peer-assessment or self-assessment) ranging on a scale from a informal, unplanned on-the-fly assessment to a formal, planned embedded assessment (Shavelson et al., 2008). In line with these ideas, Andrade (2010, p. 344) stresses that main purposes of FA are "(1) providing information about students' learning to teachers and administrators in order to guide them in designing instruction; and (2) providing feedback to students about their progress in order to help them determine how to close any gaps between their performance and the targeted learning goals". Consequently, feedback is considered being a central element of FA, which is proven being a powerful tool for supporting learning if it addresses the questions "Where am I going?", "How am I going?", and "Where to next?" (Hattie & Timperley, 2007). In their prominent works, Bennett (2011) and Hattie (2008) highlight that implementing FA into teaching can have significant effects on students' learning. That's why FA is described being the "next best hope" (Cizek, 2010, p. 2) or a "powerful tool" (Wylie et al., 2012, p. 121) in educational research and policy.

Based on this research, recently several studies have analyzed ways of implementing FA into mathematics education and of supporting mathematics teachers in doing so. Dalby and Swan (2019) underpin that using iPads can support teachers' implementation of FA in mathematics education, but that harking back to technology can cause uncertain roles of teachers, students and technology in classrooms. Rakoczy et al. (2019) point out that effects of FA on achievement and interest are not direct but indirect ones and that students' support by FA in mathematics education depends on students' perceived usefulness. In the study of Boström and Palm (2023), mathematics teachers took part in a professional development program (PDP) on FA and implemented FA into mathematics teaching afterwards. Effects on students' learning cannot be found, the authors explain this finding by a big variance of quality of implementation of FA into teaching. On the contrary, Andersson and Palm (2017) find effects of a teacher PDP on FA on students' achievement in a control trial ($d = 0.66$), but particular effects on students' specific mathematical processes do not exist. And Schütze et al. (2017) demonstrate that teachers' taking part in PDP can have indirect effects on teachers' feedback generation, but not on instructional practice in mathematics education.

Those results stress that implementing FA into teaching (of mathematics) is challenging both teachers and students and that teacher PDPs are necessary but not sufficient. This is also reported by reviews on implementation of FA highlighting the complex mechanisms of FA in classroom: Yan et al. (2021) review 52 studies exploring teachers' role in FA and point out that both contextual and personal factors influence teachers' intention and implementation regarding FA. Heitink et al. (2016) review 25 studies on implementing assessment for learning and derive a complex model of prerequisites of teachers, students and assessment all influencing each other. And Kingston and Nash (2011, p. 33) highlight in their often cited meta-analysis "that there is wide variation in the type and impact of formative assessment. Moderator analysis showed content area had the greatest impact on mean effects with mathematics".

In summary, although FA is "next best hope" in the beginning 2010, research results about ways of implementing FA and ways of supporting teachers in doing so are struggling and empirical effects are partly inconsistent.

## 3 Instructional quality of teaching (in mathematics education)

Starting from these inconsistent results and based on the idea that considering (instructional) quality of teaching is crucial when thinking about how to improve students' learning (Muijs et al., 2014; Roehring et al., 2012), some very first studies consider instructional quality of teaching for better understanding effects of the implementation of FA into teaching. Decristan et al. (2015) analyze the interplay of curriculum-embedded FA (categorized as a "specific teaching practice" by the authors) on the one hand and GIQ

(categorized as a "global factor") on the other hand. They report (interaction) effects of GIQ and FA on students' science learning. These findings are partly confirmed by Pinger et al. (2018) for mathematics education, reporting interaction effects of some aspects of GIQ (cognitive activation, classroom management) and the implementation of FA as well. By these results, both studies highlight that considering instructional quality of teaching can have an added value for better understanding effects of FA on students' learning. However, studies analyzing the added value of (interaction) effects of both GIQ and SSIQ as well as of the implementation of FA on students' (mathematics) achievement simultaneously are not known to the authors.

The current exploratory study harks back to this desideratum and analyzes the added value of considering both GIQ and SSIQ for understanding effects of the implementation of FA into teaching. GIQ and SSIQ are based on already existing (and prominent) work on instructional quality of teaching (in mathematics education) – both frameworks will be presented briefly for preparing empirical work (for further frameworks see e.g. Learning Mathematics for Teaching Project, 2011; Pianta & Hamre, 2009; Schoenfeld, 2013): (1) In line with previous work of Decristan et al. (2015) and Pinger et al. (2018) on GIQ and FA, GIQ refers to the model of "Three Basic Dimensions" (TBD; for current discussions see Praetorius et al., 2020; Praetorius & Charalambous, 2018). It is a model of GIQ being prominent (not only) in German speaking countries, being published first based on empirical findings in the context of TIMSS (Klieme et al., 2001) and being discussed in the international context theoretically based on a Swiss-German video study on teaching of Pythagoras' theorem (Klieme et al., 2009; Lipowsky et al., 2009). The model comprises three basic dimensions of GIQ (Praetorius & Charalambous, 2018, pp. 409–410): *Classroom management*, that is "identifying and strengthening desirable student behaviors, and preventing undesirable ones"; *cognitive activation*, saying that "teachers should explore and build on students' prior knowledge" and that "challenging problems and questions […] should be used to stimulate cognitive conflicts, engage students in higher-level thinking processes, and support metacognition"; and *student support*, which "primarily deals with the shared perception of the quality of social interactions". The later one is sometimes theoretically and empirically separated into *cognitive support* on the one hand and *motivational support* on the other hand (Kleickmann et al., 2020). Although studies on GIQ in mathematics education prove effects of (several) dimensions of this model on students' mathematics learning (Decristan et al., 2015; Kunter et al., 2013; Pinger et al., 2018), empirical results about the impact on students' achievement in general are inconsistent – and in many cases no effects can be reported (Praetorius & Charalambous, 2018). (2) Since the model of TBD can be discussed critically concerning the lack of subject-specifics (Praetorius & Charalambous, 2018; Schlesinger & Jentsch, 2016), based on a systematic literature survey Schlesinger et al. (2018) expand this model by two subject-specific dimensions for mathematics education. First added dimension is *subject-related quality*, e.g. comprising the mathematical depths of a lesson ("the teacher provides generalizations, mathematical connections and possibilities to structure the mathematical content") and the support of mathematical competencies ("the teacher provides the opportunity to deal with mathematical processes such as problem-solving, modelling or reasoning and proof"). Second added dimension is *teaching-related quality*, e.g. comprising the use of multiple representations and the use of mathematical examples in classroom. Both quantitative and qualitative analyses point out that this framework offers additional possibilities for understanding mathematics education as well as "the opportunity to present a more extensive and complete picture of instructional quality" (Schlesinger et al., 2018, p. 487). However, significant empirical evidence is still missing.

Despite this theoretical conceptualization and empirical verification of GIQ and SSIQ (not only of those two frameworks presented here), it must be stated (independently of the idea of understanding the implementation of FA): Effects of GIQ are inconsistent as well (Praetorius et al., 2020) and the role of SSIQ for understanding effective teaching remains complex (Charalambous & Praetorius, 2018). This is the more problematic, since subject-specific perspectives on instructional quality of teaching are claimed offering additional insights into mathematical learning processes at school (Brunner, 2018; Lindmeier & Heinze, 2020). It is imperative that further work on the added value of considering both GIQ and SSIQ when analyzing specific teaching practices (not only but also in the context of implementation of FA) is needed for a better understanding of effective teaching.

## 4 Research question and hypotheses

Based on those considerations – that is: (1) FA is a promising specific teaching practice (in mathematics education) (Bennett, 2011; Hattie, 2008), but effects of implementation are inconsistent (Kingston & Nash, 2011); (2) considering GIQ and SSIQ can have added values for understanding these inconsistent effects of implementation of FA into (mathematics) teaching (Decristan et al., 2015; Pinger et al., 2018), but effects of GIQ and SSIQ itself are inconsistent/complex as well (Charalambous & Praetorius, 2018; Praetorius et al., 2020); (3) SSIQ should be considered explicitly in further

empirical research (Brunner, 2018) –, the present exploratory study addresses the following research question:

*Research Question (RQ).* Does considering (interaction) effects of GIQ and/or SSIQ of teaching mathematics have an added value for understanding students' learning when implementing FA into mathematics education instead of solely looking at the implementation of FA itself?

*Hypotheses* can only be formulated unambiguously to a limited extent: On the one hand, Decristan et al. (2015) report (interaction) effects of dimensions of GIQ (and FA) on students' achievement, on the other hand, Pinger et al. (2018) cannot confirm these findings completely. Additionally, on the one hand empirical studies highlight the importance of considering SSIQ of teaching mathematics at least from a theoretical perspective (Charalambous & Praetorius, 2018), on the other hand, empirical evidence is rare (Lindmeier & Heinze, 2020). And empirical studies analyzing the effect of implementation of FA and SSIQ on students' learning simultaneously are not known to the authors.

## 5 Method

The current exploratory study re-analyzed data of the research project *Conditions and Consequences of Classroom Assessment*[1] ($Co^2CA$), which was funded by the German Research Foundation (GRF) within the priority program *Competence Models for Assessing Individual Learning Outcomes and Evaluating Processes.* The project lasted for six years (2007–2013) and was conducted by the German Institute for International Educational Research and the Universities of Kassel and Lueneburg. Study work finished in 2018 by the publication of Pinger et al. (2018) about an intervention study in mathematics education (lasting from October 2010 to March 2011), reporting interaction effects of GIQ (assessed by students' perception) and the implementation of FA on students' mathematics achievement. SSIQ is not considered in this or any former publication.

The study presented here reports on this intervention study as well, re-analyzing the effect of GIQ and SSIQ on students' achievement when implementing FA into every-day teaching of mathematics (focusing on the topic of Pythagoras' theorem and the competence of mathematical modelling exemplarily; see below). This is done by using newly created data both on GIQ and SSIQ being based on analyzing videotaped lessons for the first time. Students' achievements were assessed before and after the intervention. Reported information on participants, design, measures, and data analyses only refer to this special intervention.[2]

### 5.1 Participants

39 mathematics teachers/classes from 25 German middle schools (grade 9; so called "Realschule") in Hessian participated in the intervention study. All participating schools, teachers and students (as well as their parents) had been informed about the purpose of the study as well as the collection and use of data before starting the study. Written permission for videotaping classroom lessons twice during the intervention (videotaped situation 1 and 2; see below) was asked for additionally. All participants (teachers as well as students) took part in the study voluntarily, no compensation was paid.

Participating teachers/classes were assigned either to a control group (CG) or an experimental group (EG) by random. While there are videos from all 39 participating classes for the first videotaped situation, there are only 33 videos for the second videotaped situation (caused by illnesses or unexpected disruptions to the daily routine at school).

Students' outcomes at the end of the intervention are central for analyses given below. In total, 856 students participated in the achievement-test at the end of the intervention. 353 of those were female, 405 were male, 98 did not make any indication. Before starting the intervention, those students were in average 14.59 years old.

Additional information on descriptive data concerning intervention groups is given in Table 1.

### 5.2 Design

The intervention covered 13 consecutive lessons (each lasting for 45 min), introducing into Pythagoras' theorem (lessons 1 to 5) and focusing on mathematical modelling in lessons 6 to 13 exemplarily[3]. Pythagoras' theorem had not been taught in participating classes before. All participating teachers from CG and EG were told to implement a unit structure of teaching consisting of four phases (see Fig. 1): (Phase 1) Introduction of Pythagoras' theorem including at least one proof and some technical tasks for training, focusing primarily on building up a general understanding

[1] The project was supported by grants from the German Research Foundation (KL 1057/10 – 3, BL 275/16 – 3, LE 2619/1–3); principal researchers: E. Klieme, K. Rakoczy (both Frankfurt), W. Blum (Kassel), D. Leiss (Lueneburg).

[2] The authors would like to thank the principal researchers of the $Co^2CA$ project for allowing them to analyze the videotaped lessons and to link them to the performance data of the students in the context of this article.

[3] Both the focus on mathematical content and mathematical competence are chosen for practical reasons: Pythagoras' theorem as well as mathematical modelling are mandatory in mathematics education in Germany, additionally extensive preparatory work of principal researchers concerning Pythagoras' theorem (Klieme et al., 2009) and mathematical modelling (Blum, 2011) exists.

**Table 1** Descriptive data concerning intervention groups

| | | | Videos | | Students | | |
|---|---|---|---|---|---|---|---|
| | $N_{Sch}$ | $N_{T/C}$ | $N(1)_{Vid}$ | $N(2)_{Vid}$ | $N_{Stud}$ | $A_{Stud}$ | $G_{Stud}$ |
| **CG + EG** | 25 | 39 | 39 | 33 | 856 | 14.59 (0.67) | 353/405/98 |
| **CG** | 11 | 15 | 15 | 14 | 331 | 14.57 (0.63) | 137/155/39 |
| **EG** | 14 | 24 | 24 | 19 | 525 | 14.60 (0.69) | 216/250/59 |

$N_{Sch}$: Number of participating schools

$N_{T/C}$: Number of participating teachers/classes

$N(X)_{Vid}$: Number of classes with videotaped lessons at videotaped situation X = 1 or X = 2

$N_{Stud}$: Number of participating students in the achievement-test at the end of the intervention

$A_{Stud}$: Aggregated students' age at 30.09.2010 (before starting intervention); standard deviation in brackets

$G_{Stud}$: Number of students giving an indication of gender (female/male/no indication)

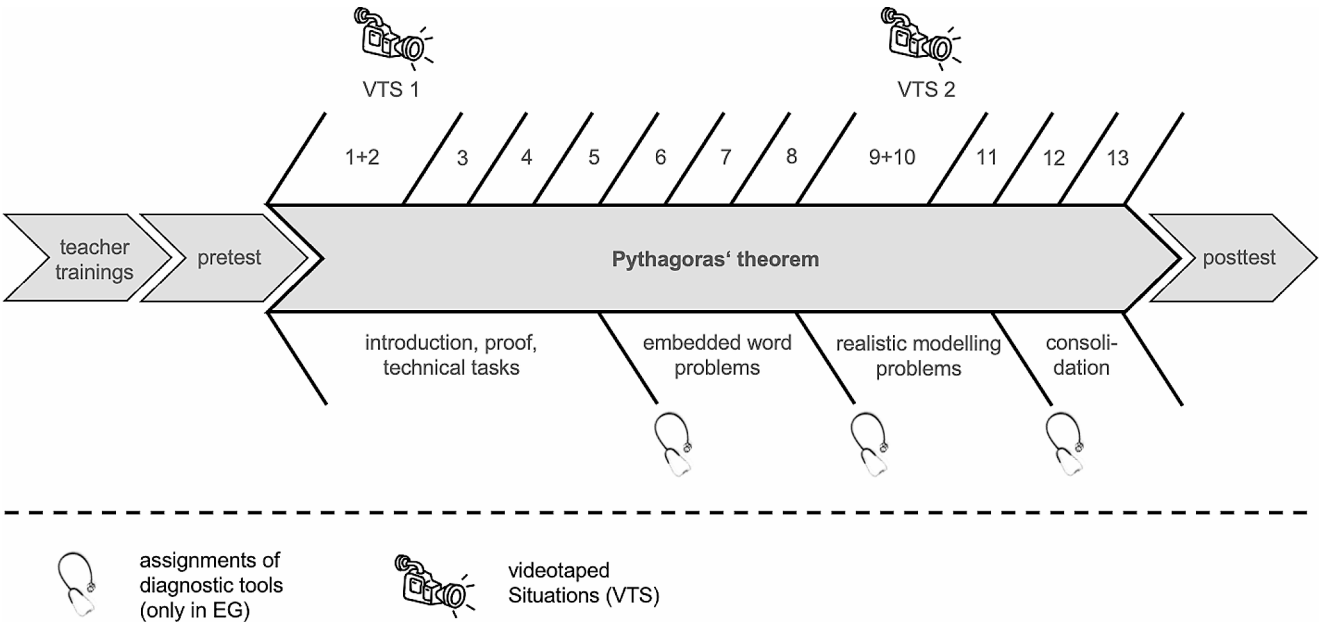
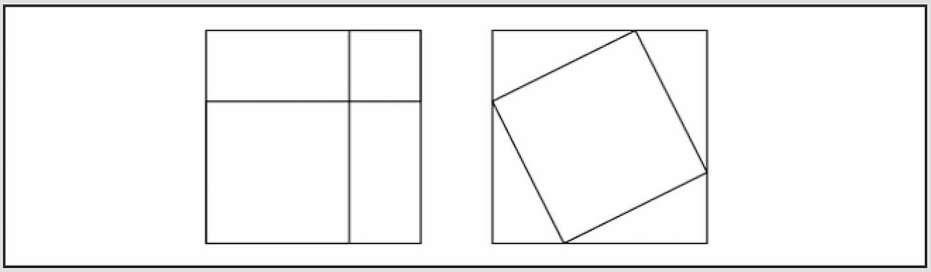



**Fig. 1** Design of the intervention study (adapted from Pinger et al., 2016)

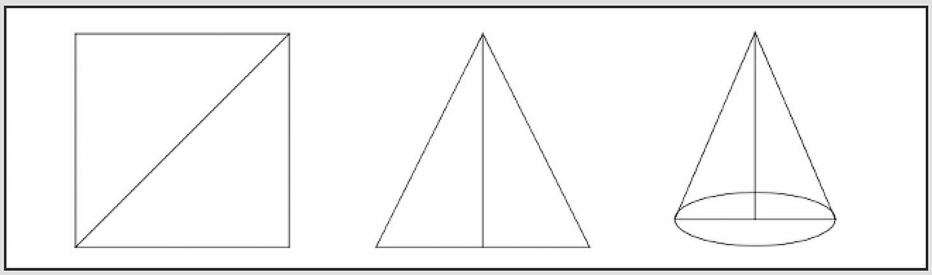
of central elements of Pythagoras' theorem (in line with Drollinger-Vetter, 2011). (Phase 2) Students had to work on embedded word problems (Maaß, 2010) for making them to use Pythagoras' theorem in simple real world situations. (Phase 3) More realistic (but not necessarily authentic) mathematical modelling problems covering the whole modelling cycle (Blum & Leiss, 2007; Kaiser, 2020) were implemented into mathematics teaching for supporting students in building up modelling competencies when dealing with Pythagoras' theorem. (Phase 4) Additional lessons were used for consolidating former learning processes and for offering the opportunity to reflect ones' own competencies. A concrete schedule, some possible proofs as well as mandatory and voluntary tasks had been given to the teachers before starting intervention. A sample proof and some sample tasks are given in Fig. 2.

All teachers from CG and EG had to participate in a first teacher training before starting the intervention (for details see Besser et al., 2015). The teachers were informed about content and processes of the study, teaching material was discussed, and the predesigned teaching unit was recapitalized. Additionally, those teachers from EG had to join a second teacher training before starting the intervention. Based on ideas about FA (Black & Wiliam, 2009; Hattie & Timperley, 2007), those teachers were trained implementing a diagnostic tool assessing students' performances regularly (at the end of lesson 5, 8 and 11) and giving task-related, process-oriented, individual written feedback to each student at the beginning of the next lesson (beginning of lessons 6, 9 and 12). This diagnostic tool (see Fig. 3 for an example) consisted of one or two tasks students had to work on and some feedback about personal strengths (processes that had been mastered), weaknesses (processes that need further improvements) as well as hints/strategies on how to continue (for more details see Pinger et al., 2016).

**Mandatorily:** Teachers were asked to use the following two sketches to proof Pythagoras‘ theorem in classroom.

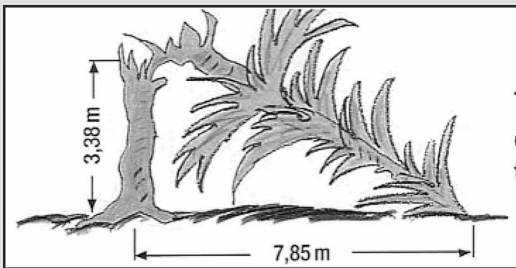


**Voluntarily:** Teachers were able to use the following sketches for making students’ identifying rectangle triangles, naming triangular sides and formulating Pythagoras’ theorem.



**PHASE 1**
**(lessons 1 to 5)**
*introduction*
*proof*
*technical tasks*

**Voluntarily:** Teachers were able to administer the following embedded word problem to their students.




The tree was snapped off during a storm. How long is the broken crown of the tree?

**PHASE 2**
**(lessons 6 to 8)**
*embedded*
*word problems*

**Mandatorily:** Teachers were asked to administer the following realistic modelling problem to their students.

The 30.7 m high "Roter Sand" lighthouse was built in the Bay of Bremen directly on the coast in 1884. Its purpose was to warn ships by its beacon that they were approaching the coast. How far away from the coast was a ship when it saw the lighthouse for the first time?

**PHASE 3**
**(lessons 9 to 11)**
*realistic*
*modelling problems*

**Fig. 2** Sample proof and sample tasks (technical tasks, embedded word problem, realistic modelling problem)

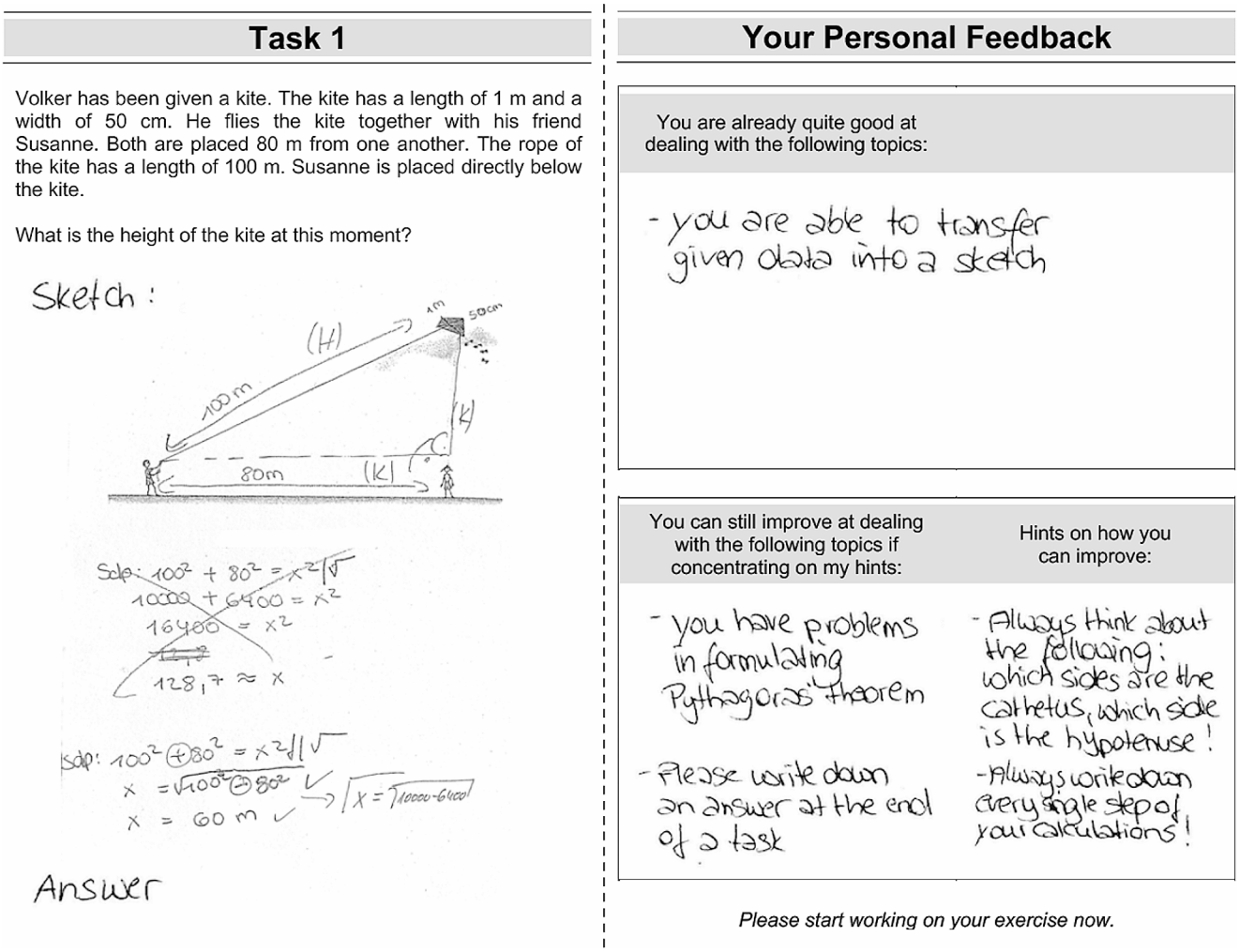



**Fig. 3** Example of the diagnostic tool at the end of lesson 8; embedded word problem on the left, teachers' feedback on the right (see also Pinger et al., 2016)

Assessments of students' mathematics achievement took place immediately before and after the teaching unit. GIQ and SSIQ is based on videotapes recording lesson 1 and 2 (videotaped situation 1; dealing with an introduction and a proof of Pythagoras' theorem) as well as lesson 9 and 10 (videotaped situation 2; dealing with mathematical modelling problems). Those lessons were held as double lessons.

### 5.3 Measures

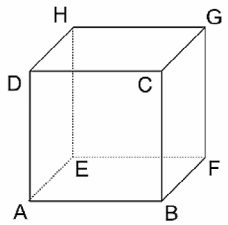
*Achievement* (A1 and A2; student level). Students' achievement was assessed by a pretest before starting the intervention (achievement 1; A1) and by a posttest at the end of the intervention (achievement 2; A2). Since students worked on Pythagoras' theorem in the intervention for the first time, the pretest asked for students' prior knowledge and consisted of 19 items dealing with mathematical content being necessary for handling Pythagoras' theorem (e.g.: using variables and simplifying terms, extracting the square root of numbers, identifying a right angle and right-angled triangles and different sides of a triangle). The posttest consisted of 17 items and assessed students' capability to use Pythagoras' theorem successfully when working on tasks focusing on technical problems, embedded word problems and realistic modelling problems. Both tests lasted for 45 min. Items were administered as a paper-pencil-test, item formats were single choice response or open response (sample items are given in Fig. 4). All items were coded dichotomously (0 = incorrect; 1 = correct) by three trained raters, interrater-reliability was very good (about 10% of students' answers were double-coded by all raters; two-way mixed (absolute) intraclass correlation (*ICC(3,3)*) was calculated both for pretest and posttest; see Table 2). Students' test-scores were one-dimensional Rasch-scaled. Since all items had been analyzed previously in a scaling study ($N = 1570$, Harks et al., 2014), item-parameters were fixed on those study results for scaling. *Weighted likelihood estimators were used as students' achievement scores* for pretest (A1) and posttest (A2).

**PRETEST (19 items)**

**Sample items 1 and 2 (prior knowledge):** Students must simplify a term and extract a square root.

| Determine the value for x (x > 0): | Calculate the following length: |
|---|---|
| $271 = 46 + x^2$ | $\sqrt{(21cm)^2-(5cm)^2}$ |

**Sample item 3 (prior knowledge):** Students must identify rectangular triangles in an inner-mathematical context.

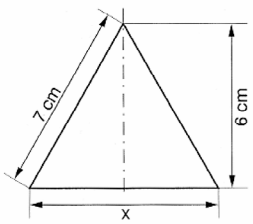



Given is the cube on the left. Tick all triangles within this cube that are right-angled:

- ☐ △ ABC
- ☐ △ BEG
- ☐ △ AFG
- ☐ △ AFC
- ☐ △ HEB

**POSTTEST (17 items)**

**Sample item 4 (technical problem):** Students must work on a technical problem by using Pythagoras' theorem.




Calculate the missing side length x in the on the left (drawing not to scale).

**Sample item 5 (realistic modelling problem):** Students must work on a realistic modelling problem by using Pythagoras' theorem.

Every year in Bad Dinkelsdorf on 1 May, the traditional dance takes place around the so-called maypole, an approx. 8 m high tree trunk. The dancers hold ribbons in their hands that are attached to the top of the maypole. With these 15 m long ribbons they dance around the maypole in such a way that in the course of the dance a beautiful pattern is created on the trunk (in the photo such a pattern can already be seen at the top of the maypole).

At what distance from the maypole do the dancers stand at the beginning of the dance (the ribbons are taut)? Describe your solution.

**Fig. 4** Sample items of pretest and posttest

**Table 2** Overview of measures

| Measures | Instr. | # Items | ICC |
|---|---|---|---|
| **Student/Class Level** | | | |
| *A1*: Achievement 1 | Pretest | 19 | .99 |
| *A2*: Achievement 2 | Posttest | 17 | .99 |
| **Class Level** | | | |
| *FA*: Implementation of Formative Assessment (CG or EG) | ---- | 1 | --- |
| *GIQ 1a*: Generic Instructional Quality 1a – Classroom Management | VTS 1 | 7 | .80 |
| *GIQ 1b*: Generic Instructional Quality 1b – Cognitive Activation | VTS 1 | 8 | .77 |
| *GIQ 1c*: Generic Instructional Quality 1c – Cognitive Support | VTS 1 | 9 | .76 |
| *GIQ 2a*: Generic Instructional Quality 2a – Classroom Management | VTS 2 | 7 | .78 |
| *GIQ 2b*: Generic Instructional Quality 2b – Cognitive Activation | VTS 2 | 8 | .57 |
| *GIQ 2c*: Generic Instructional Quality 2c – Student Support | VTS 2 | 9 | .77 |
| *SSIQ 1*: Subject-Specific Instructional Quality 1 – Pythagoras' Theorem | VTS 1 | 12 | .89 |
| *SSIQ 2*: Subject-Specific Instructional Quality 2 – Mathematical Modelling | VTS 2 | 10 | .68 |

ICC: For interpretation of ICC see Cicchetti (1994)

VTS: Videotaped Situation

*EAP/PV*-Reliability was 0.66 for the pretest and 0.74 for the posttest. *ICC(1)* indicating the proportion of total variance which can be attributed to between-differences is 0.15 for pretest and 0.09 for posttest.

*Implementation of FA* (FA; class level). For controlling for potential differences due to the group allocation, the implementation of FA is assessed by a *dummy-coded variable*, either indicating that students/classes are assigned to CG (FA = 0) or to EG (FA = 1).

*Generic instructional quality* (GIQ 1a-c and GIQ 2a-c; class level). Based on the model of Three Basic Dimensions (Praetorius et al., 2020), all videos from the videotaped situations 1 and 2 were analyzed regarding GIQ by using high-inference ratings for describing classroom management, cognitive activation and cognitive support: The rating scale for *classroom management* (GIQ 1a and GIQ 2a) consisted of seven items, e.g. asking for classroom disruptions, the teacher's monitoring and clarity of rules. The rating scale for *cognitive activation* (GIQ 1b and GIQ 2b) consisted of eight items, e.g. asking for the insistence on reasons and explanations, the support for cognitive autonomy and the implementation of challenging problems. *Student support* (GIC 1c and GIQ 2c) was assessed by cognitive support (and not motivational support; see above), since empirical evidence indicates effects of cognitive support (and not of motivational support) on students' achievement (Kleickmann et al., 2020). The rating scale consisted of nine items, e.g. asking for the clarity of goals, the coherence of teaching and the supportive use of feedback. All scales had already been used in previous studies (Lotz et al., 2013) and have only been adapted for study use slightly.

Three trained raters (university pre-service teacher students; trainings lasted for several days using sample video material) analyzed the generic quality of teaching (*ICC(3,3)* is good to very good, see Table 2), the rating of the items ranged from 1 (appearing nearly never) to 4 (appearing extremely often), the mean score of all items was the scale score. *The mean scores of all three raters were used as scores for GIQ* of classroom management, cognitive activation, and student support – separated for videotaped situation 1 (GIQ 1a-c) and videotaped situation 2 (GIQ 2a-c).

*Subject-specific instructional quality (SSIQ 1 and SSIQ 2; class level).* Based on the work of Schlesinger et al. (2018), GIQ is extended by elements of SSIQ by assessing mathematical depth and support of mathematical competence as part of subject-related quality. In lesson 1 and 2 (videotaped situation 1), students were introduced to Pythagoras' theorem by working on a proof and some technical tasks. It was analyzed, whether "the teacher provided generalizations, mathematical connections and possibilities to structure the mathematical content" by an appropriate thematization of so called "elements of understanding of Pythagoras' theorem" (SSIQ 1). The scale was adapted from Drollinger-Vetter and Lipowsky (2006) and Drollinger-Vetter (2011) and consisted of twelve items, e.g. asking for the thematization of different types of triangular sides in rectangular triangles as well as the thematization and combination of algebraic and/or geometric formulations of Pythagoras' theorem. In lesson 9 and 10 (videotaped situation 2), students worked on realistic modelling problems which can be solved by using Pythagoras' theorem in real world contexts. It was analyzed, whether "the teacher provided the opportunity to deal with mathematical processes" (SSIQ 2) by demanding students to conduct central steps of mathematical modelling (Kaiser, 2020). The scale covered those processes of mathematical modelling which are being described by a well-established modelling cycle (Blum & Leiss, 2007), which are justified by a activity-theoretical analysis of mathematical modelling (Böhm, 2013), and which are already being used by many empirical case studies to analyze the

quality of implementation of mathematical modelling into teaching (for an overview see Vorhölter et al., 2019). The scale consisted of ten items, e.g. asking for the appropriate implementation of having to understand given information, of having to make assumptions, of having to work on the modelling problem mathematically as well as of having to interpret and to validate mathematical results.

Items of both scales were scored dichotomously (0 = no thematization in classroom; 1 = appropriate thematization in classroom) by three trained raters (*ICC(3,3)* is good to very good, see Table 2). For scoring 1, the content covered by an item had to be thematized substantially for mor than just a few seconds (as beeing standard in many subject-specific frameworks, see e.g. Learning Mathematics for Teaching Project, 2011). Scale-score was the sum of item-scores, normed on a range from zero to one (for an easier interpretation; scores indicate the percentage of thematized content). *For describing SSIQ, the mean score of all three raters was used.*

An overview of measures is given in Table 2, additional information on instruments assessing GIQ and SSIQ is given in an electronic supplement (ES01). Since measures of achievement are used both on the student and the class level within analyses (see below), those measures are listed as "student/class level".

### 5.4 Data analysis

Data analyses for the preparation of descriptive data were conducted by using SPSS 27. For doing so, student data (*A1* and *A2*) was aggregated on a class level, while class level data was used as given in Table 2.

Data analyses focusing on the research question were conducted by using Mplus 7 (Muthén & Muthén, 2012). Achievement 1 (*A1*) and Achievement 2 (*A2*) were z-standardized on the student level (level 1; within), variables assessing instructional quality (*GIQ* and *SSIQ*) were z-standardized on the class level (level 2, between). The implementation of FA (*FA*) was implemented as dummy (0 = CG; 1 = EG) on the class level. Multilevel regression analyses were conducted, several models were compared: In all models, students' achievement in the posttest (*A2*) was dependent variable, students' achievement in the pretest (*A1*; both on student and class level) and the implementation of FA (*FA*; on class level) were independent variables. In selected models, instructional quality of teaching (*GIQ* and *SSIQ*) as well as interactions of FA and instructional quality (*FA*GIQ* and *FA*SSIQ*) were additional independent variables.

Caused by relatively few cases on the class level, missing data was dealt with by listwise exclusion of cases.

## 6 Results

Descriptive statistics including sample size, mean scores, standard deviations, and correlations are given in the electronical supplement (ES02). It can be summarized: Firstly, the implementation of FA *(FA)*, that is belonging to CG or EG, is not correlated with any measure of instructional quality. Instructional quality is neither systematically better nor worse if teachers/classes are advised to FA. Secondly, subdimensions of *GIQ* correlate significantly among each other ($0.34 < r < .76$) on a medium level. Additionally, GIQ of videotaped situation 1 (*GIQ 1a-c*) does not correlate with SSIQ of videotaped situation 1 (*SSIQ 1*), but GIQ of videotaped situation 2 (*GIC 2a-c*) correlates with SSIQ of videotaped situation 2 (*SSIQ 2*) significantly.

Regarding the research question, selected standardized model results are given in Table 3 and 4 (only models with interaction effects are reported here; all models are given in an electronic supplement (ES03)). Firstly, across different models it can be stated that it is mainly students' achievement in the pretest *(A1)* predicting students' achievement in the posttest, whereas effects of implementation of FA *(FA)* cannot be found. Secondly, effects of *GIQ* ($\beta = 0.78$) as well as interaction effects ($\beta$=-*0.71*) reducing the main effect can be found for classroom management in videotaped situation 1 (*GIQ 1a*; model 10), but not in videotaped situation 2 (*GIQ 2a*; model 13). Thirdly, some effects of SSIQ are given (model 08, 09, 16 and 17). *SSIQ 1* (Pythagoras' theorem in videotaped situation 1) is predicting students' achievement in the posttest positively ($\beta = 0.28$; model 08), the effect even becomes bigger when considering the interaction effect of *FA* and *SSIQ 1* ($\beta = 1.06$; model 16). Interaction effect in this model is negative, that is high quality of *SSIQ* reduces the influence of *FA* on students' achievement. Additionally, and on the contrary, *SSIQ 2* (mathematical modelling in videotaped situation 2) is predicting students' achievement in the posttest negatively ($\beta$=-*0.41* and $\beta$=-*0.60*; model 09 and model 17).

## 7 Discussion

The present exploratory study aims at analyzing the added value of considering (interaction) effects of GIQ and SSIQ on students' mathematics achievement when implementing FA into mathematics teaching. *Results* reveal some important findings: Firstly, there are no direct effects of implementation of FA on students' learning. It can be stated once again (see also Pinger et al., 2018) that changing classroom teaching (by implementing FA) does not have to imply improving classroom teaching immediately. Although teachers of the current study were trained to implement

**Table 3** Multilevel regression analyses predicting students' achievement in the posttest (A2) – part 1 (all models are given in ES03)

| Variable | Model 10 | | | Model 11 | | | Model 12 | | | Model 13 | | | Model 14 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | β | SE | *p* | β | SE | *p* | β | SE | *p* | β | SE | *p* | β | SE | *p* |
| **Student Level** | | | | | | | | | | | | | | | |
| A1: Achievement 1 | **.42** | .03 | .00 | **.42** | .03 | .00 | **.42** | .03 | .00 | **.43** | .04 | .00 | **.43** | .04 | .00 |
| **Class Level** | | | | | | | | | | | | | | | |
| A1: Achievement 1 | **.56** | .14 | .00 | **.64** | .15 | .00 | **.65** | .14 | .00 | **.73** | .13 | .00 | **.66** | .14 | .00 |
| FA: Implementation of FA (Dummy Coded; 0 = CG; 1 = EG) | -.32 | .16 | .04 | .22 | .16 | .18 | .22 | .17 | .20 | .05 | .18 | .78 | .07 | .17 | .70 |
| GIQ 1a: Generic Instructional Quality 1a Classroom Management | **.78** | .24 | .00 | | | | | | | | | | | | |
| GIQ 1b: Generic Instructional Quality 1b Cognitive Activation | | | | .06 | .28 | .83 | | | | | | | | | |
| GIQ 1c: Generic Instructional Quality 1c Cognitive Support | | | | | | | .42 | .29 | .16 | | | | | | |
| GIQ 2a: Generic Instructional Quality 2a Classroom Management | | | | | | | | | | -.19 | .37 | .62 | | | |
| GIQ 2b: Generic Instructional Quality 2b Cognitive Activation | | | | | | | | | | | | | -.25 | .27 | .35 |
| GIQ 2c: Generic Instructional Quality 2c Student Support | | | | | | | | | | | | | | | |
| SSIQ 1: Subject-Specific Instructional Quality 1 – Pythagoras' Theorem | | | | | | | | | | | | | | | |
| SSIQ 2: Subject-Specific Instructional Quality 2 – Mathematical Modelling | | | | | | | | | | | | | | | |
| Interaction (FA*GIQ or FA*SSIQ) | **-.71** | .20 | .00 | -.15 | .26 | .56 | -.36 | .24 | .13 | .02 | .31 | .96 | .23 | .21 | .29 |
| | ***R²*** | ***SE*** | ***p*** | ***R²*** | ***SE*** | ***p*** | ***R²*** | ***SE*** | ***p*** | ***R²*** | ***SE*** | ***p*** | ***R²*** | ***SE*** | ***p*** |
| **R-Square (Student Level)** | **.18** | .03 | .00 | **.18** | .03 | .00 | **.18** | .03 | .00 | **.19** | .03 | .00 | **.19** | .03 | .00 |
| **R-Square (Class Level)** | **.65** | .12 | .00 | **.43** | .20 | .03 | **.29** | .18 | .01 | **.56** | .19 | .00 | **.57** | .20 | .01 |

**Table 4** Multilevel regression analyses predicting students' achievement in the posttest (A2) – part 2 (all models are given in ES03)

| Variable | Model 15 | | | Model 16 | | | Model 17 | | |
|---|---|---|---|---|---|---|---|---|---|
| | β | SE | *p* | β | SE | *p* | β | SE | *p* |
| **Student Level** | | | | | | | | | |
| A1: Achievement 1 | **.43** | .04 | .00 | **.43** | .04 | .00 | **.43** | .04 | .00 |
| **Class Level** | | | | | | | | | |
| A1: Achievement 1 | **.70** | .16 | .00 | **.53** | .13 | .00 | **.72** | .12 | .00 |
| FA: Implementation of FA (Dummy Coded; 0 = CG; 1 = EG) | .07 | .18 | .68 | .04 | .17 | .82 | .07 | .15 | .67 |
| GIQ 1a: Generic Instructional Quality 1a Classroom Management | | | | | | | | | |
| GIQ 1b: Generic Instructional Quality 1b Cognitive Activation | | | | | | | | | |
| GIQ 1c: Generic Instructional Quality 1c Cognitive Support | | | | | | | | | |
| GIQ 2a: Generic Instructional Quality 2a Classroom Management | | | | | | | | | |
| GIQ 2b: Generic Instructional Quality 2b Cognitive Activation | | | | | | | | | |
| GIQ 2c: Generic Instructional Quality 2c Student Support | -.06 | .35 | .87 | | | | | | |
| SSIQ 1: Subject-Specific Instructional Quality 1 – Pythagoras' Theorem | | | | **1.06** | .41 | .01 | | | |
| SSIQ 2: Subject-Specific Instructional Quality 2 – Mathematical Modelling | | | | | | | **-.60** | .18 | .00 |
| Interaction (FA*GIQ or FA*SSIQ) | -.04 | .27 | .88 | **-.83** | .39 | .03 | .24 | .15 | .12 |
| | $R^2$ | *SE* | *p* | $R^2$ | *SE* | *p* | $R^2$ | *SE* | *p* |
| **R-Square (Student Level)** | **.19** | .03 | .00 | **.18** | .03 | .00 | **.19** | .03 | .00 |
| **R-Square (Class Level)** | **.54** | .20 | .00 | **.60** | .15 | .00 | **.74** | .18 | .00 |

FA explicitly, effective implementation of FA is challenging. This is in line with research on teachers' professional development, pointing out that changing teaching practices successfully needs supporting teachers in doing so over a long period of time (Darling-Hammond et al., 2017). Secondly, these interpretations are additionally underpinned by (interaction) effects of *GIQ 1a* (concretely: classroom management), which are given immediately after starting to implement FA into teaching mathematics (model 10) but are disappearing a few lessons later (model 13), although correlations of *GIQ 1a* and *GIQ 1b* are at least medium ones ($r = .67$). Positive effects of classroom management ($\beta = 0.78$), which had also been found by Lipowsky et al. (2009), are counterbalanced by negative interaction effects with *FA* ($\beta = -0.71$), so classroom management has nearly a zero effect size in EG ($\beta = 0.78–0.71 = 0.07$). Or in other words: Implementing FA is extremely complex (Kingston & Nash, 2011; Yan et al., 2021), teachers' routines seem to break down. Thirdly, referring to negative effects of SSIQ on students' learning (models 09 and 17), it is imperative to understand that changing teachers' way of teaching mathematics on a content level by implementing a specific teaching practices (here: FA) also might cause negative effects on students' learning, at least in the short term (which is in line with theory about teachers' expertise and about deliberate practice, see e.g. Berliner, 1995; Ericsson et al., 1993). In this special case, teachers had to implement mathematical modelling into teaching mathematics – it is well known, that this is challenging for teachers (Blum, 2015).

All those findings stress the added value of considering GIQ and SSIQ when analyzing the impact of implementation of FA on learning, However, *implications* for research must be discussed carefully in the context of the operationalization of GIQ and SSIQ and its interplay with FA: According to Decristan et al. (2015), in the present study GIQ and SSIQ are operationalized being "global factors of instructional quality", while FA is operationalized being a "specific teaching practice" (see also Good et al., 2009). However, GIQ/SSIQ on the one hand and FA on the other hand are not necessarily disjoint constructs, on the contrary "formative assessment may be closely related to global factors" (Decristan et al., 2015, p. 1138). This is obvious for GIQ immediately (e. g.: "exploring and building on students' prior knowledge" is a key concept of FA as well), but also relevant for conceptualizations of SSIQ: For example the subject specific framework "TRU Math" (Schoenfeld, 2013, 2014) comprises the "use of assessment" ("building on students' ideas") being an inherent component of high quality of teaching mathematics. Future studies should focus this interplay of (subject specific) instructional quality of teaching (mathematics) and specific teaching practices for further improving understanding effective teaching.

Some design-based and methodological *limitations* must be addressed when interpreting the results: (a) *Design.* As has already been mentioned, the teacher training for implementing FA into teaching Pythagoras' theorem only lasted for several hours. This is caused by careful and conscious considerations about supporting teachers to implement central elements of teaching on the one hand and about not making teachers to overtime while taking part in the study on the other hand. However, effective teacher PDPs should at best last for several weeks/months, not for several hours (Darling-Hammond et al., 2017). (b) *Measures I.* GIQ and SSIQ are assessed by using instruments having been slightly adapted from prior empirical studies (Drollinger-Vetter & Lipowsky, 2006 (SSIQ 1); Lotz et al., 2013 (GIQ)) or having been newly developed based on theoretical considerations and empirical work of many case studies (Vorhölter et al., 2019 (SSIQ 2). For GIQ, this decision is based on the empirically-based idea of assessing students' support by cognitive support and not by motivational support (Kleickmann et al., 2020). For SSIQ, this has been done because the framework offered by Schlesinger et al. (2018) is conceptualized for on-the-fly assessment of nine different subdimensions of SSIQ, not allowing deeper analyses of single subject-specific sub-dimensions (here: mathematical depth of the lesson; support of mathematical competencies) of quality of instruction. Since implementation of "elements of understanding concerning Pythagoras' theorem" and the implementation of "mathematical modelling" is focused, those conceptualizations are used. (c) *Measures II.* Assessments of GIQ and SSIQ are based on videotaped classroom observations of external (trained) observers. Although those ratings are usually much more objective than those from students or teachers, open questions of validity (of scoring, generalization, extrapolation and implication) must be thought about consciously (Bell et al., 2012; Praetorius et al., 2014). However, in general validity of those external observations is high (for a brief discussion see Schlesinger & Jentsch, 2016). (d) *Measures III*. SSIQ is assessed by low-inference ratings. While low-inference ratings are generally more reliable than high-inference ratings, those ratings are quite often used to describe the quantity/frequency of events, not its quality. However, in this special case SSIQ is operationalized by analyzing whether "the teacher provides a specific opportunity to learn", which is in line with theory (see above) and which is also done by prominent studies describing mathematics instructional quality (Learning Mathematics for Teaching Project, 2011; National Center for Education Statistics, 2003). (e) *Data analyses*. Since data analyses are based on measures of instructional quality on the class level, the sample size consists of only 39 or 33 classes on level 2 respectively. This is unproblematic for interpreting both regression coefficients and variance

components, but standard errors might be estimated too small in some cases (Maas & Hox, 2005). Furthermore, since participating classes are distributed equally between CG and EG (concerning age, gender), only prior knowledge is used as independent variable on the student level, results must be interpreted with this limitation. (g) *Topicality of the study*. Data being used is based on an intervention study from 2010/2011. Analyses possibly do not describe the "current state of teaching in German mathematics classroom"– which is of course not intended by this study. But based on design, measures, and analyses, reported results support understanding the complexity of implementing FA into teaching mathematics by considering both GIQ and SSIQ.

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1007/s11858-024-01562-2.

**Funding** Open Access funding enabled and organized by Projekt DEAL.

## Declarations

**Conflict of interest** No potential conflict of interest is reported by the authors.



## References

Andersson, C., & Palm, T. (2017). The impact of formative assessment on student achievement: A study of the effects of changes to classroom practice after a comprehensive professional development programme. *Learning and Instruction*, *49*, 92–102. https://doi.org/10.1016/j.learninstruc.2016.12.006.

Andrade, H. L. (2010). Summing up and moving forward: Key challenges and future directions for research and development in formative assessment. In H. L. Andrade, & G. J. Cizek (Eds.), *Handbook of formative assessment* (pp. 344–352). Routledge.

Bell, C. A., Gitomer, D. H., McCaffrey, D. F., Hamre, B. K., Pianta, R. C., & Qi, Y. (2012). An argument approach to observation protocol validity. *Educational Assessment*, *17*(2–3), 62–87. https://doi.org/10.1080/10627197.2012.715014.

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

Berliner, D. C. (1995). Teacher expertise. In C. J. Anderson (Ed.), *International encyclopedia of teaching and teacher education* (pp. 46–52). Elsevier.

Besser, M., Leiss, D., & Klieme, E. (2015). Wirkung von Lehrerfortbildungen auf Expertise von Lehrkräften zu Formativem Assessment im kompetenzorientierten Mathematikunterricht. *Journal of Developmental and Educational Psychology*, *47*(2), 110–122. https://doi.org/10.1026/0049-8637/a000128

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education*, *5*(1), 7–74.

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

Blömeke, S., Gustafsson, J. E., & Shavelson, R. J. (2015). Beyond dichotomies: Competence viewed as a continuum. *Zeitschrift für Psychologie*, *223*(1), 3–13. https://doi.org/10.1027/2151-2604/a000194.

Blum, W. (2011). Can modelling be taught and learnt? Some answers from empirical research. In G. Kaiser, W. Blum, R. Borromeo Ferri, & G. Stillmann (Eds.), *Trends in teaching and learning of mathematical modelling* (pp. 15–30). Springer.

Blum, W. (2015). Quality teaching of mathematical modelling: What do we know, what can we do? In S. J. Cho (Ed.), *The proceedings of the 12th international congress on mathematical education. Intellectual and attitudinal challenges* (pp. 73–96). Springer. https://doi.org/10.1007/978-3-319-12688-3_9.

Blum, W., & Leiss, D. (2007). How do students and teachers deal with modelling problems? In C. Haines, P. Galbraith, W. Blum, & S. Khan (Eds.), *Mathematical modelling (ICTMA 12): Education, engineering and economics* (pp. 222–231). Horwood. https://doi.org/10.1533/9780857099419.5.221.

Böhm, U. (2013). *Modellierungskompetenzen langfristig und kumulativ fördern. Tätigkeitstheoretische Analyse des mathematischen Modellierens in der Sekundarstufe I*. Springer Spektrum. https://doi.org/10.1007/978-3-658-01821-4.

Boström, E., & Palm, T. (2023). The effect of a formative assessment practice on student achievement in mathematics. *Frontiers in Education*, *8*, 1101192. https://doi.org/10.3389/feduc.2023.1101192.

Brunner, E. (2018). Qualität von Mathematikunterricht: Eine Frage der Perspektive. *Journal Für Mathematik-Didaktik*, *39*, 257–284. https://doi.org/10.1007/s13138-017-0122-z.

Charalambous, C. Y., & Praetorius, A. K. (2018). Studying mathematics instruction through different lenses: Setting the ground for understanding instructional quality more comprehensively. *ZDM – Mathematics Education*, *50*, 355–366. https://doi.org/10.1007/s11858-018-0914-8.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*(4), 284–290.

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.

Dalby, D., & Swan, M. (2019). Using digital technology to enhance formative assessment in mathematics classrooms. *British Journal of Educational Technology*, *50*(2), 832–845. https://doi.org/10.1111/bjet.12606.

Darling-Hammond, L., Hyler, M. E., & Gardner, M. (2017). *Effective teacher professional development*. Learning Policy Institute.

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: How do they interact in promoting science understanding? *American Educational Research Journal*, *52*(6), 1133–1159. https://doi.org/10.3102/0002831215596412.

Drollinger-Vetter, B. (2011). *Verstehenselemente und strukturelle Klarheit. Fachdidaktische Qualität der Anleitung von mathematischen Verstehensprozessen im Unterricht*. Waxmann.

Drollinger-Vetter, B., & Lipowsky, F. (2006). Fachdidaktische Qualität der Theoriephasen. In E. Klieme, C. Pauli, & K. Reusser (Eds.), *Dokumentation der Erhebungs- und Auswertungsinstrumente zur schweizerisch-deutschen Videostudie. Unterrichtsqualität, Lernverhalten und mathematisches Verständnis. Teil 3. Videoanalysen* (pp. 189–205). Materialien zur Bildungsforschung.

Ericsson, K. A., Krampe, R. T., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review*, *100*(3), 363–406.

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

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

Hattie, J. (2008). *Visible learning. A synthesis of over 800 meta-analyses relating to achievement*. Routledge. https://doi.org/10.4324/9780203887332.

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

Heitink, M. C., Van der Kleij, F. M., Veldkamp, B. P., Schildkamp, K., & Kippers, W. B. (2016). A systenatic review of prerequisites for implementing assessment for learning in classroom practice. *Educational Research Review*, *17*, 50–62. https://doi.org/10.1016/j.edurev.2015.12.002.

Johnson, C. C., Sondergeld, T. A., & Walton, J. B. (2019). A study of the implementation of formative assessment in three large urban disrticts. *American Educational Research Journal*, *56*(6), 2408–2438. https://doi.org/10.3102/0002831219842347. *56*.

Kaiser, G. (2020). Mathematical modelling and applications in education. In S. Lerman (Ed.), *Encyclopedia of mathematics education* (pp. 553–561). Springer. https://doi.org/10.1007/978-3-030-15789-0_101.

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

Kleickmann, T., Steffensky, M., & Praetorius, A. K. (2020). Quality of teaching in science education. More than three basic dimensions? In A.-K. Praetorius, J. Grünkörn, & E. Klieme (Eds.), *Empirische Forschung zu Unterrichtsqualität. Theoretische Grundfragen und quantitative Modellierungen* (pp. 37–55). Beltz Juventa. https://doi.org/10.25656/01:25862.

Klieme, E., Schümer, G., & Knoll, S. (2001). Mathematikunterricht in der Sekundarstufe I: Aufgabenkultur und Unterrichtsgestaltung. In E. Klieme, & J. Baumert (Eds.), *TIMSS-Impulse für Schule und Unterricht. Forschungsbefunde, Reforminitiativen, Praxisberichte und Video-Dokumente* (pp. 43–57). Bundesministerium für Bildung und Forschung.

Klieme, E., Pauli, C., & Reusser, K. (2009). The pythagoras study: Investigating effects of teaching and learnign in Swiss and German mathemathics classrooms. In T. Janik, & T. Seidel (Eds.), *The power of video studies in investigating teaching and learning in the classroom* (pp. 137–160). Waxmann.

Kunter, M., Klusmann, U., Baumert, J., Richter, D., Voss, T., & Hachfeld, A. (2013). Professional competence of teachers: Effects on instructional quality and student development. *Journal of Educational Psychology*, *105*(3), 805–820. https://doi.org/10.1037/a0032583.

Learning Mathematics for Teaching Project. (2011). Measuring the mathematical quality of instruction. *Journal of Mathematics Teacher Education*, *14*, 25–47. https://doi.org/10.1007/s10857-010-9140-1.

Lindmeier, A., & Heinze, A. (2020). Die fachdidaktische Perspektive in der Unterrichtsqualitätsforschung: (bisher) ignoriert, implizit enthalten oder nicht relevant? In A.-K. Praetorius, J. Grünkorn, & E. Klieme (Eds.), *Empirische Forschung zu Unterrichtsqualität. Theoretische Grundfragen und quantitative Modellierungen* (pp. 255–268). Beltz Juventa. https://doi.org/10.25656/01:25878.

Lipowsky, F., Rajoczy, K., Drolliner-Vetter, B., Klieme, E., Reusser, K., & Pauli, C. (2009). Quality of geometry instruction and its short-term impact on students' underrstanding of pythagorean theorem. *Learning and Instruction*, *19*(6), 527–537. https://doi.org/10.1016/j.learninstruc.2008.11.001.

Lotz, M., Lipowsky, F., & Faust, G. (2013). Technischer Bericht zu den PERLE-Videostudien. In F. Lipowsky, & G. Faust (Eds.), *Dokumentation der Erhebungsinstrumente des Projekts "Persönlichkeits- und Lernentwicklung von Grundschulkindern" (PERLE) – Teil 3*. Materialien zur Bildungsforschung.

Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology*, *1*(3), 86–92. https://doi.org/10.1027/1614-2241.1.3.86.

Maaß, K. (2010). Classification scheme for modelling tasks. *Journal für Mathematik-Didaktik*, *31*(2), 285–311. https://doi.org/10.1007/s13138-010-0010-2.

Muijs, D., Kyriakides, L., Van der Werf, G., Creemers, B., Timperley, H., & Earl, L. (2014). State of the art – teacher effectiveness and professional learning. *School Effectiveness and School Improvement*, *25*(2), 231–256. https://doi.org/10.1080/09243453.2014.885451.

Muthén, L. K., & Muthén, B. O. (2012). *Mplus user's guide. Seventh edition*.

National Center for Education Statistics (2003). *Teaching mathematics in seven countries. Results from the TIMSS 1999 video study*.

Patrick, H., Mantzicopoulos, P., & Sears, D. (2012). Effective classrooms. In K. R. Harris, S. Graham, & T. Urdan (Eds.), *APA educational psychology handbook. Volume 2: Individual differences and cultural and contextual factors* (pp. 443–469). American Psychological Associtaion. https://doi.org/10.1037/13274-020.

Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher*, *38*(2), 109–119. https://doi.org/10.3102/0013189X09332374.

Pinger, P., Rakoczy, K., Besser, M., & Klieme, E. (2016). Implementation of formative assessment – effects of quality of programme delivery on students' mathematics achievement and interest. *Assessment in Education: Principles Policy & Practice*, *25*(2), 160–182. https://doi.org/10.1080/0969594X.2016.1170665.

Pinger, P., Rakoczy, K., Besser, M., & Klieme, E. (2018). Interplay of formative assessment and instructional quality – interactive effects on students' mathematics achievement. *Learning Environmental Research*, *21*, 61–79. https://doi.org/10.1007/s10984-017-9240-2.

Praetorius, A. K., & Charalambous, C. Y. (2018). Classroom observation frameworks for studying instructional quality: Looking back and looking forward. *ZDM – Mathematics Education*, *50*, 535–553. https://doi.org/10.1007/s11858-018-0946-0.

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.

Praetorius, A. K., Klieme, E., Kleickmann, T., Brunner, E., Lindmeier, A., Taut, S., & Charalambous, C. Y. (2020). Towards developing a theory of generic teaching quality. Origin, current status, and necessary next steps regarding the three basic dimensions model. In A.-K. Praetorius, J. Grünkorn, & E. Klieme (Eds.), *Empirische Forschung zu Unterrichtsqualität. Theoretische Grundfragen*

*und quantitative Modellierungen* (pp. 15–36). Beltz Juventa. https://doi.org/10.25656/01:25861.

Rakoczy, K., Pinger, P., Hochweber, J., Klieme, E., Schütze, B., & Besser, M. (2019). Formative assessment in mathematics: Mediated by feedback's perceived usefulness and students' self-efficacy. *Learning and Instruction*, *60*, 154–165. https://doi.org/10.1016/j.learninstruc.2018.01.004.

Roehring, A. D., Turner, J. E., Arrastia, M. C., Christesen, E., McElhaney, S., & Jakiel, L. M. (2012). Effective teachers and teaching: Characteristics and practices related to positive student outcomes. In K. R. Harris, S. Graham, & T. Urdan (Eds.), *APA educational psychology handbook. Volume 2: Individual differences and cultural and contextual factors* (pp. 501–527). American Psychological Associtaion. https://doi.org/10.1037/13274-020.

Schlesinger, L., & Jentsch, A. (2016). Theoretical and methodological challenges in measuring instructional quality in mathematics education using classroom observations. *ZDM – Mathematics Education*, *48*, 29–40. https://doi.org/10.1007/s11858-016-0765-0.

Schlesinger, L., Jentsch, A., Kaiser, G., König, J., & Blömeke, S. (2018). Subject-specific characteristics of instructional quality in mathematics education. *ZDM – Mathematics Education*, *50*, 475–490. https://doi.org/10.1007/s11858-018-0917-5.

Schoenfeld, A. H. (2013). Classroom observations in theory and practice. *ZDM – Mathematics Education*, *45*, 607–621. https://doi.org/10.1007/s11858-012-0483-1.

Schoenfeld, A. H. (2014). What makes for powerful classrooms, and how can we support teachers in creating them? *Educational Researcher*, *43*(8), 404–412. https://doi.org/10.3102/0013189X14554450.

Schütze, B., Rakoczy, K., Klieme, E., Besser, M., & Leiss, D. (2017). Training effects on teachers' feedback practice: The mediating function of feedback knowledge and the moderating role of self-efficacy. *ZDM – Mathematics Education*, *49*, 475–489. https://doi.org/10.1007/s11858-017-0855-7.

Seidel, T., & Shavelson, R. J. (2007). Teaching effectiveness research in the past decade: The role of theory and research design in disentangeling meta-analysis results. *Review of Educational Research*, *77*(4), 454–499. https://doi.org/10.3102/0034654307310317.

Shavelson, R. J., Young, D. B., Ayala, C. C., Brandon, P. R., Furtak, E. M., & Ruiz-Primo, M. A. (2008). On the impact of curriculum-embedded formative assessment on learning: A collaboration between curriculum and assessment developers. *Applied Measurement in Education*, *21*(4), 295–314. https://doi.org/10.1080/08957340802347647.

Vorhölter, K., Greefrath, G., Ferri, B., Leiss, R., D., & Schukajlow, S. (2019). Mathematical modelling. In H. N. Jahnke & L. Hefendehl-Hebeker (Eds.), *Traditions in German-speaking mathematics education research* (pp. 91–114). Springer. https://doi.org/10.1007/978-3-030-11069-7_4.

Wylie, E., Gullickson, A., Cummings, K., Egelson, P., Noakes, L., & Norman, K. (2012). *Improving formative assessment practice to empower student learning*. Corwin.

Yan, Z., Li, Z., Panadero, E., Yang, M., Yang, L., & Lao, H. (2021). A systematic review on factors influencing teachers' intentions and implementations regarding formative assessment. *Assessment in Education: Principles Policy & Practice*, *28*(3), 228–260. https://doi.org/10.1080/0969594X.2021.1884042.