Zurich Open Repository and Archive
University of Zurich
University Library
Strickhofstrasse 39
CH-8057 Zurich
www.zora.uzh.ch

Year: 2024

Capturing the subject-specific quality of mathematics instruction: how do expert judgments relate to students' assessments of the quality of their own learning and understanding?

Pauli, Christine ; Lipowsky, Frank ; Reusser, Kurt

DOI: https://doi.org/10.1007/s11858-024-01561-3

Posted at the Zurich Open Repository and Archive, University of Zurich
ZORA URL: https://doi.org/10.5167/uzh-261444
Journal Article
Published Version

The following work is licensed under a Creative Commons: Attribution 4.0 International (CC BY 4.0) License.

Originally published at:
Pauli, Christine; Lipowsky, Frank; Reusser, Kurt (2024). Capturing the subject-specific quality of mathematics instruction: how do expert judgments relate to students' assessments of the quality of their own learning and understanding? ZDM - Mathematics Education, 56(5):893-905.
DOI: https://doi.org/10.1007/s11858-024-01561-3

ZDM – Mathematics Education
https://doi.org/10.1007/s11858-024-01561-3

ORIGINAL PAPER

Capturing the subject-specific quality of mathematics instruction: How do expert judgments relate to students' assessments of the quality of their own learning and understanding?

Christine Pauli[1] · **Frank Lipowsky**[2] · **Kurt Reusser**[3]

Accepted: 6 March 2024
© The Author(s) 2024

Abstract
Based on an opportunity-use model of instructional quality, this study investigates the extent to which subject-specific instructional quality rated by experts is reflected in students' assessments of their own learning and understanding, and how students' perceptions predict their achievement. The analyses used data from a German-Swiss sample of 36 classes with around 900 lower secondary students, obtained as part of the so-called "Pythagoras study" in the school year 2002/2003. The teachers were instructed to introduce the Pythagorean theorem in three lessons, which were videotaped. Using the videos, the experts assessed the instruction quality with respect to the goal of promoting a deep understanding of the theorem. The students completed the questionnaires assessing their understanding of the content, their learning process, and the general comprehension orientation of the teacher. The results showed significant and moderate correlations on the class level between expert-rated subject-specific teaching quality and students' perceptions of their own learning and understanding, as well as of the teacher's general comprehension orientation. Multilevel models revealed that subject-specific expert ratings are reflected in individual students' perceptions of their own learning and understanding. Student perceptions were also associated with achievement gains. The results suggest that the assessment of quality by students and experts is more closely linked if a distinction is made between the quality of the learning opportunities offered and their use and if subject-specific criteria are used instead of generic criteria. This study contributes to a more nuanced understanding of the validity of student perspective in assessing instructional quality.

Keywords Subject-specific teaching quality · Students' perceptions of their own learning process · Mathematics instruction · Opportunity-use models

1 Introduction

Cognitive-constructivist theories of learning emphasize students' active role in learning with respect to the goal of deep and robust understanding (Brunner, 2018; Chi & Wylie, 2014; Koedinger et al., 2012; Reusser, 2006). This is also reflected in "opportunity-use-models" of teaching and learning (Fend, 1998; Helmke, 2003; Reusser & Pauli, 2010; Vieluf & Klieme, 2023). According to these models, students' learning processes cannot be controlled from outside; rather, their learning depends on them making the most effective use of the learning opportunities offered by the teacher. This suggests that student perspective should also be considered when examining teaching quality. Students' perceptions have been included in various studies on teaching quality and, in some cases, compared with teacher or expert ratings (e.g. Cheng et al., 2023; Clausen, 2002; Fauth et al., 2020; Göllner et al., 2021; Kunter & Baumert, 2006; Wang & Eccles, 2016). Most of these studies have revealed surprisingly low correlations between expert and student perspectives. Apparently, students are rather poor at assessing deep structural features of instructional quality.

However, there are two limitations to the abovementioned research. First, the students were mostly asked about their perceptions of the learning opportunities and not about their use of these opportunities. We assume that students are quite able to assess important aspects of instructional

✉ Christine Pauli
christine.pauli@unifr.ch

[1] University of Fribourg, Fribourg, Switzerland

[2] University of Kassel, Kassel, Germany

[3] University of Zurich, Zurich, Switzerland

Published online: 20 March 2024

quality when asked about their own learning and understanding processes (Jansen et al., 2022; Merk et al., 2021). Second, the expert and student ratings used in these studies were mostly related to the generic aspects of teaching quality, although subject-specific aspects seem to be particularly important for promoting a deep and robust understanding of concepts (cf. Drollinger-Vetter, 2011; Hiebert & Grouws, 2007; Schlesinger & Jentsch, 2016; Schlesinger et al., 2018; Schoenfeld, 2018).

The analyses presented here address both aspects. We investigate the question of how expert and student judgments are related when expert judgments refer to subject-specific quality characteristics of teaching and student judgments refer to the use of these learning opportunities, as well as the relation between student judgments and learning progress. The present analyses draw on the data and earlier analyses in the context of the so-called "Pythagoras study" (Klieme et al., 2009; Rakoczy et al., 2007). The database generated in this study facilitates a comparison of perspectives, considering both the opportunity-use model and subject-specific quality characteristics, even though the data was already obtained in 2002 (see 6.2 for a discussion).

2 Theoretical framework

2.1 Teaching toward understanding–a crucial feature of instructional quality

Enabling and fostering deep conceptual understanding is a central goal for all students not only in mathematics instruction but in all school subjects (Gravemeijer et al., 2017; Patrick et al., 2012; Reusser & Reusser-Weyeneth, 1997). Understanding content means mindfully realizing the inner relations of concepts, i.e. recognizing the connections between and within concepts, and using them for generating new ideas and thoughts (e.g. Aebli, 1980/1981; Drollinger-Vetter, 2011; Koedinger et al., 2012; Wertheimer, 1945). From a cognitive-constructivist perspective, the development of thoroughly understood mathematical knowledge by students requires a focused and intensive cognitive processing of the learning contents in terms of deep learning. This includes using prior knowledge, establishing connections between existing and new knowledge, constructing mental models, and integrating newly created knowledge into one's own knowledge base through constructive cognitive activities (Hiebert & Carpenter, 1992; Schlesinger & Jentsch, 2016).

For comprehension-oriented mathematics teaching, the challenge is to initiate, instruct, and support students in these processes. It has become increasingly clear that comprehension-orientated instruction depends less on specific methods than on deep structural features, such as the extent to which students' cognitive engagement is stimulated and supported as they engage with the learning content (Hiebert & Grouws, 2007; Mayer, 2009; Praetorius & Charalambous, 2018; Prediger et al., 2022; Reusser, 2005).

2.2 From generic to subject-specific concepts of teaching quality: challenges of measurement with a focus on comprehension

In recent years, the deep structural quality of teaching has often been measured using a three-dimensional framework. It is based on the three dimensions of cognitive activation, student support, and classroom management (Klieme et al., 2009; Kunter & Voss, 2013), which are also referred to as "three basic dimensions" (TBDs) of teaching quality. Both observation protocols (Lipowsky et al., 2009; Praetorius et al., 2018) and student questionnaires (e.g. Herbert et al., 2022; Senden et al., 2023) were developed on the basis of the TBDs. The "cognitive activation" dimension is particularly important for assessing comprehension orientation, as it focuses on the extent and quality of students' cognitive activities. Observation protocols record, for example, the quality of the tasks being worked on or certain characteristics of classroom interactions (cf. Praetorius et al., 2018).

Most of these instruments capture "cognitive activation" as a generic feature of instructional quality that does not focus on specific subjects or content. However, there is a growing consensus that teaching quality also includes subject-specific aspects. It is not a matter of creating as *many* connections as possible but of recognizing and elaborating the elements and relationships that are *relevant* for the understanding of a concept in a transparent and coherent development process. Comprehension-oriented mathematics instruction helps learners focus on essential principles and conceptual elements, and links them through active, in-depth processing into correct and coherent subject-specific knowledge and thought structures (e.g. Hiebert & Grouws, 2007; Prediger et al., 2022). For (research on) teaching, this demands a subject-specific analysis of the content-related pedagogical requirements for understanding a particular concept or content structure (Drollinger-Vetter, 2011; Schlesinger & Jentsch, 2016). In fact, the call for greater consideration of the subject-specific perspective on teaching quality has gained prominence in recent years (e.g. Brunner, 2018; Dreher & Leuders, 2021; Lindmeier & Heinze, 2020; Schlesinger et al., 2018).

2.2.1 Measuring instructional quality with consideration for opportunity-use models

Recently, various observation protocols have been developed that aim to capture the subject-related pedagogical aspects of the quality of mathematics lessons (for comprehensive

overviews, see e.g., Charalambous & Praetorius, 2018; Schlesinger & Jentsch, 2016). What these instruments have in common is that their application requires a high level of expertise in mathematics as well as in the didactics of mathematics (Dreher & Leuders, 2021). Therefore, it is not surprising that studies on the subject-specific quality of mathematics teaching have so far relied primarily on expert ratings.

Based on opportunity-use models that view teaching quality as an interplay between the provision and use of learning opportunities (Vieluf & Klieme, 2023), the assessment should not only focus on the teacher's providing behavior but also include the students' use of it. It can be assumed that observers' judgments of quality not only refer to features of teachers' tasks and behavior but are also likely to include the perceived features of student behavior to some extent (Fauth et al., 2020). Thus, when assessing cognitive activation, it can be assumed that the signs of students' cognitive activity or attention are partially considered, and conclusions are drawn about the quality of students' understanding and learning processes. Global assessments, however, only partially capture how individual students use the offering based on their learning prerequisites. One possible solution is to include student perceptions.

2.2.2 Measuring instructional quality from students' viewpoints

For years, it has been a common approach in classroom research to measure instruction and its quality from students' viewpoints (e.g. De Jong & Westerhof, 2001; Wagner et al., 2013). However, only a few studies have systematically differentiated between the perception of the learning opportunities offered (e.g., tasks and teacher behavior) and their use (e.g., cognitive engagement and comprehension).

The validity of student ratings of instructional quality has been extensively studied (e.g. Clausen, 2002; Herbert et al., 2022; Kunter & Baumert, 2006; Lenske & Praetorius, 2020; Scherer et al., 2016; Senden et al., 2023; Wagner et al., 2013; Wisniewski et al., 2020). Overall, the results on the *discriminant validity* of such student judgments show that students can discriminate among different dimensions of teaching quality, including those of the TBDs (e.g. Fauth et al., 2014; Senden et al., 2023). Students' perceptions of teaching quality have certain *prognostic validity* for cognitive, motivational, and social student outcomes. Overall, the empirically observed associations between students' perceptions of teaching quality and achievement gains are rather small and concern classroom management rather than cognitive activation (Herbert et al., 2022; Kunter & Baumert, 2006; Scherer et al., 2016; Senden et al., 2023; Waldis et al., 2010; Wallace et al., 2016). In terms of *convergent validity*, comparisons of student perceptions with expert or teacher ratings show significant relations if the ratings refer to well-observable characteristics of instruction and to the social aspects of learning, but–with a few exceptions (e.g. Cheng et al., 2023)–no or less pronounced agreement for characteristics such as cognitive activation and learning support (e.g. De Jong & Westerhof, 2001; Kunter & Baumert, 2006).

As a possible explanation for the rather low correlations between student and observer judgments, Lenske and Praetorius (2020) showed that the students sometimes misunderstood the items of an assessment questionnaire. This was especially the case for "cognitive activation." In addition, rating scales that focus on instructional quality sometimes contain a mix of items that ask about perceptions of instructional features and items that focus on student behaviors (see also Fauth et al., 2020).

2.2.3 Revisiting the measurement of instructional quality from students' perspectives as a desideratum of research

Note that the student questionnaires used in most of the above studies did not record teaching quality in a subject-specific manner and were primarily aimed at the perception of learning opportunities and not their use. Even though several instruments for observer ratings of subject-specific quality characteristics have been developed recently (see 2.2.1), only a few studies have surveyed both expert and student perceptions in connection with subject-specific quality features of comprehension-oriented teaching (e.g. Cheng et al., 2023; Scherer & Gustafsson, 2015). This is understandable, as an assessment of subject-specific quality characteristics requires subject-specific and didactic expertise. Based on an opportunity-use model, it makes sense to ask students about the *use* of learning opportunities and not about the subject-specific characteristics of these opportunities. The latter can be better assessed by experts. Students' perceptions of their own understanding and learning can show how they use the offer (Jansen et al., 2022; Merk et al., 2021; Rieser & Decristan, 2023; Vieluf, 2022).

So far, little is known about how closely students' assessments of their own learning and comprehension processes are related to expert assessments that focus on subject-specific features of comprehension-oriented mathematics instruction. Earlier analyses in the context of the Pythagoras study, which addressed a different question and included only one aspect of subject-specific teaching quality, have already indicated such a relation (Rakoczy et al., 2007). Using the same database, the present study examines the extent to which experts' subject-specific quality judgments are reflected in students' perceptions of their own learning and understanding and the teacher's general orientation

toward understanding, as well as the extent to which students' self-assessments predict their learning gains. The aim is to contribute to the investigation of the convergent and prognostic validity of student perceptions considering both subject-and content-related characteristics of teaching quality and an opportunity-use model.

3 Research questions

The following research questions (RQs) are addressed:

1a) Are expert ratings of subject-specific quality features of instruction related to students' (class-level) perceptions of their own learning process and content understanding when both students and experts refer to the same mathematics lessons?
1b) Are expert ratings of subject-specific quality features of instruction related to students' (class-level) ratings of the teacher's overall comprehension orientation (as measured at the end of the school year)?
2a) Are expert ratings of subject-specific quality features of instruction related to individual students' perceptions of their own learning process and content understanding when both relate to the same mathematics lessons?
2b) Are expert ratings of subject-specific quality features of instruction related to individual student perceptions of the teacher's overall comprehension orientation?
3) Are individual student perceptions of their own learning and understanding and of the teacher's overall understanding orientation related to learning success when central learning prerequisites are controlled for?

4 Method

4.1 Database

The present study is embedded in the quasi-experimental project "Quality of instruction, learning, and mathematical understanding" (also called the "Pythagoras study"), which investigated the impact of mathematics instruction on students' cognitive and motivational outcomes over the period of one school year (Hugener et al., 2009; Klieme et al., 2009; Lipowsky et al., 2009). In particular, the present analysis extends previous analyses that investigated the cognitive and motivational effects of structured instruction (Rakoczy et al., 2007). The original sample comprised 20 Swiss and 20 German classes with 1015 students from two lower secondary school types: the highest track (*Gymnasium)* and the intermediate track (*Realschule, Sekundarschule*).[1] Participation was voluntary. The present analyses draw on data from a reduced sample comprising 36 classes[2] and a maximum of 913 students. In all classes, a three-lesson unit on the "Introduction to Pythagorean Theorem" and a two-lesson unit on algebraic word problems were videotaped in the school year 2002/2003.

4.2 Study design

The analyses presented are based on the "Pythagorean unit." The teachers were asked to submit at least one proof (of any kind) of the Pythagorean theorem; otherwise, they were free to organize the lessons as they wished. Data were collected at four measurement time points. At the beginning of the school year, the students' general mathematics achievement and cognitive ability were tested. In addition, their characteristics such as interest in mathematics were assessed using a questionnaire (T1). Immediately before the three-lesson unit "Introduction to the Pythagorean Theorem," the students were tested on their prior knowledge of the Pythagorean theorem (T2). Immediately after the three lessons, the students assessed their learning and understanding (questionnaire, T3). Subsequently, learning success was assessed (post-test Pythagoras, T3). At the end of the school year, the students' overall mathematics achievement was tested again, and their general perceptions of mathematics instruction and the mathematics teacher were recorded (T4).

4.3 Measures of students' perceptions

The scale items can be found in the Appendix online ("Supplementary Information"). The instruments used were developed as part of the Pythagoras study.

Scale "students' perceptions of their learning process" The students' self-reported cognitive processes during the three Pythagoras lessons were assessed at T3 using four items (4-point-response scale, items: see Appendix). The reliability coefficient (Cronbach's alpha) was 0.86, the mean of the scale was 3.34, and the standard deviation was 0.58 (Rakoczy et al., 2005).

[1] Since the Pythagorean theorem is part of the ninth-grade curriculum in Germany and the eighth-grade curriculum in Switzerland, we included German ninth-grade classes and eighth-grade classes from the German-speaking part of Switzerland.

[2] Four classes were excluded from the analyses (2 teachers cancelled the study before it was completed; in one class, the students' data were incomplete; and in another class, the teacher did not teach the content as required).

Single item "students' perception of their own attained Pythagoras-related understanding" The students' perceptions of their own attained understanding were assessed at T3 using the single item "How well have you understood the content that you went through?" The students responded to this question on a six-point scale (see Appendix). The mean of the item was 5.07, and the standard deviation was 0.97 (Rakoczy et al., 2005). The higher the value, the higher the students rated their comprehension.

Scale "students' perceptions of the overall comprehension orientation of the math teacher" The students' perceptions of the comprehension orientation of the teacher and his/her teaching were assessed at T4 (end of the school year) using a bipolar 4-item scale. The students had to decide which pole best represented their opinion (six-point response scale, see Appendix). The higher the value, the higher the students estimated the teacher's focus on understanding. The reliability was $\alpha = 0.86$, and the mean of the scale was $M = 4.67$ ($SD = 1.08$).

Correlations were found among the three measures (Table A1, Appendix). The association was stronger when both perceptions referred to the Pythagorean unit ($r = 0.66^{**}$). In contrast, the correlations were lower if one of the two variables related to the teachers' overall comprehension orientation and thus to the long-term period of the school year ($r = 0.22^{**}$; $r = 0.11^{**}$).

Calculation of ICC(1) and ICC(2) Beyond the individual perceptions of the students, aggregated student judgments can be used as a source of information about the comprehension orientation of the lessons and the teacher as characteristics of the shared environment. For this, it is necessary to calculate *ICC*(1) and *ICC*(2) (Lüdtke et al., 2009). The results for *ICC*(1) revealed that between 9 and 15% of the variance in student perceptions could be explained by belonging to the class; consequently, there were substantial differences between the classes. The values for *ICC*(2) lied above 0.70 or just below (Table A2, Appendix). Therefore, it can be assumed that student judgments reflect differences between the learning conditions in classes and not idiosyncratic perceptions.

4.4 Measures of students' achievements, mathematics-related interest, and general cognitive abilities

Mathematical achievement The data from all achievement tests were scaled using the ConQuest program (Wu et al., 1997), based on a one-parameter item-response model (Rasch model; for details see Lipowsky et al., 2006). The four achievement tests used were scaled independently of each other.

At T1, we measured the general mathematics knowledge with 10 items regarding basic skills, understanding mathematical proofs, and application ability. The EAP/PV reliability was 0.60, and the weighted mean-square residuals (MSQs) were between 0.96 and 1.08.

The students' mathematics achievement before and after the Pythagorean theorem was measured in a content-specific manner. The Pythagorean pre-test (T2) focused on the major prerequisites for a conceptual understanding of the Pythagorean theorem. Individual achievement scores were estimated separately for the pre-test and post-test using item response theory. For the pre-test, the mean square parameters of the items ranged between 0.92 and 1.03; the EAP/PV reliability was 0.64.

The Pythagorean post-test (T3) focused on the conceptual understanding of the Pythagorean theorem and its application to simple tasks. The post-test took additional 15 min to complete. The mean square parameters of the items ranged between 0.89 and 1.24; the EAP/PV reliability was 0.78.

At T4, we measured the general knowledge with 18 items referring to basic skills, algebra skills, and application ability. The EAP/PV reliability score was 0.72, and the weighted MSQs were between 0.91 and 1.09.

Scale "mathematics-related interest" Interest in mathematics was recorded at T1 using 8 items in the student questionnaire (Appendix). Answers were recorded on a four-point scale. The reliability coefficient (Cronbach's alpha) was 0.91, the mean of the scale was 2.69, and the standard deviation was 0.71 (Lipowsky et al., 2009; Rakoczy et al., 2005).

General cognitive ability test The students' general cognitive abilities were measured at T1 using a subtest of the Heller and Perleth (2000) cognitive abilities test. The mean amounted to 50.98 points, and the standard deviation was 9.96 points. This is in line with the mean of 50 points and the standard deviation of 10 points for the *t*-scale (Lipowsky et al., 2009).

4.5 Expert ratings of subject-specific instructional quality

Two experts–a co-author and an expert in mathematics education–assessed the subject-specific quality of the videotaped lessons using a rating protocol that they had jointly developed under the leadership of the mathematics education expert. They focused on the theory and proof phases of the lessons in which concepts were introduced, theorems were stated, and proofs were given (Drollinger-Vetter, 2011, p. 227; Drollinger-Vetter et al., 2006, Appendix). To comprehensively assess the subject-specific quality, they focused on the occurrence of conceptual elements ("elements of understanding": EoU), the quality of modes of representation, and

the structural clarity of the content covered in the Pythagoras lessons (Drollinger-Vetter & Lipowsky, 2006).

Score "occurrence of EoU" The experts analyzed the content frame of the Pythagorean theorem by asking what conceptual elements (EoU) of the Pythagorean theorem a teacher should address in the introductory instruction to promote students' sustained understanding of this content (Drollinger-Vetter, 2011). Nine EoU were identified as essential for a deeper understanding of the Pythagorean theorem (see Appendix). The experts rated whether these EoU were treated in the three-lesson unit and, if so, whether it was extensive or short. Each EoU was scored with 1 (=no occurrence), 2 (=short occurrence), or 3 (=extensive occurrence).

The occurrence of the EoU was rated independently by two experts (Drollinger-Vetter & Lipowsky, 2006). This procedure led to sufficient rater reliability for six of the nine items. For these six items, the generalizability coefficient (G) for relative decisions amounted over 0.65. For one of the nine items, the coefficient was lower than 0.58. For this reason, they brought about a consensus decision for those elements on which they had not reached an agreement. Because the ratings of the nine EoU must not necessarily correlate with each other, they computed the sum score of the occurrence of the nine elements. The mean of the sum scores was $M=22.45$ ($SD=3.51$), and the range varied between 14 and 27.

Scale "quality of modes of representation" The EoU can be treated in different representational modes. The quality of each of the four modes of representation (formal, verbal, iconic, and enactive) was assessed separately on a four-point scale ranging from 1 (=low) to 4 (=high). For each class, an overall assessment was given. The inter-rater reliability was assessed by computing the value of G for relative decisions for any of the four items. The coefficients ranged between 0.66 and 0.85. The mean of the scale was $M=2.84$ ($SD=0.60$) and Cronbach's alpha was $\alpha=0.73$.

Scale "structural clarity of content" The scale "structural clarity" was based on four items (Appendix), which were rated between 1 (low) and 4 (high). The inter-rater reliability was assessed by computing the value of G for relative decisions for each of the four items. The coefficients ranged between 0.72 and 0.84. The mean of the scale was $M=2.70$ ($SD=0.65$) and Cronbach's alpha was $\alpha=0.88$.

Interrelations of the three dimensions Since the content of the three subject-specific dimensions overlapped and the coding of the EoU formed the basis for the other two dimensions, "quality of modes of representation" and "structural clarity of content", the three dimensions might be inter-related. In fact, the correlations were high and significant (0.74**, 0.75**, 0.83**, Appendix, Table A3). An exploratory factor analysis (principal component analysis) showed that the three dimensions loaded on one factor. This superordinate factor can explain 84.6% of the total variance of all three items. The reliability of the scale comprising the three dimensions was $\alpha=0.89$. For further analyses, we used the mean of this scale, which we called "subject-specific quality of the Pythagorean unit." It represents the occurrence, the quality, and the linking of EoU and the forms of representation.

4.6 Analysis methods

For RQs 1a and b, the expert ratings were correlated with the aggregated perceptions of the students at the class level. To exclude the possibility of the correlations being influenced by class composition, we controlled for the mean mathematics achievement (T1) and the mean mathematics-related interest (T1) of the class.

To analyze the relation between the subject-specific quality of the Pythagorean unit rated by the experts and the students' *individual* perceptions of their own learning process and comprehension and the general comprehension orientation of the teacher (RQs 2a, b), we used multilevel analyses. In addition, the prediction of achievement gains by individual student perceptions was analyzed through multilevel modeling (RQ 3) using HLM 6.04 (Raudenbush et al., 1993). All variables included in the analyses were grand-mean centered on level 1 (student) and grand-mean centered on level 2 (class) (cf. Lüdtke et al., 2009). Based on previous research (cf. Herbert et al., 2022), the students' individual perceptions were assumed to be influenced by student characteristics (e.g., ability and interest) as well as factors on class level (e.g., mathematics performance of the class). Therefore, the mean achievement of the class in the Pythagoras pretest (T2) was included as a control variable in Models 1 and 2 (Table 2). For Model 3 (Table 2), which was used to analyze the long period of the whole school year, the mean achievement in the general mathematics test (T1) was controlled for. At the student level, we controlled for mathematics achievement (T2 or T1), interest in mathematics, and general cognitive ability.

5 Results

Regarding RQs 1a and b, Table 1 shows that the raters' assessments on subject-specific instructional quality were moderately aligned with the students' (class-level) perceptions of their own learning processes, their own comprehension of the content as recorded immediately after the

Table 1 Correlations between subject-specific quality of instruction (observer ratings) and students' perceptions (partial correlations controlling for mathematical achievement and math-related interest, T1)

	Students: own learning process (Pyth)	Students: own understanding (Pyth)	Students: teacher's overall comprehension orientation (T4)
Expert ratings: Quality (Pyth)	0.51**	0.47**	0.39*

$N = 36$ classes ** $p < .01$

Pythagoras unit but also with the perception of the general comprehension orientation of the mathematics teacher, which was recorded at the end of the school year.

For RQs 2a and b, we analyzed whether individual student perceptions can be predicted by the expert ratings of subject-specific quality of instruction (Table 2). In Models 1 and 2, the focus was on students' perceived learning and understanding related to the three Pythagoras lessons, which were also rated by the experts (RQ 2a). From a long-term perspective, the focus in Model 3 was on the individual perception of the teacher's comprehension orientation in general, which was recorded at the end of the school year (RQ 2b). According to the results, the expert ratings of subject-specific instructional quality are reflected in the differences in students' individual perceptions of their own learning processes and understanding after controlling for student and class characteristics. The regression weights for the subject-specific instructional quality are $\beta = 0.16$ in Model 1 and $\beta = 0.12$ in Model 2. For the overall understanding orientation of the teacher, this relation only emerged as a trend ($\beta = 0.20$, $p < 0.10$).

Models 1 and 2 indicate that the students' mathematics-related interest and individual mathematics performance are important predictors for their perceptions, while class performance has a negative effect. Model 1 explained 16.74% of the variance in students' perceptions of their learning process; in Model 2, the included variables explained 20.23% of the variance in students' perceptions of their own comprehension.

Regarding RQ 3, Models 4 to 6 (Table 3) were used to examine the extent to which individual student perceptions can predict the development of mathematics achievement in the Pythagoras teaching unit or over the entire school year. This is the case for the perception of one's own learning and understanding as well as the overall comprehension orientation of the teacher when controlling for other important influencing factors. However, the regression weights $\beta = 0.09$, $\beta = 0.11$, and $\beta = 0.06$ are low.

6 Discussion

The starting point for our study were findings from previously conducted teaching quality research, which showed that the convergent and prognostic validity of student assessments of cognitive activation was rather limited. This concerns a quality dimension that is expected to play a central role in comprehension-oriented teaching. In this research, cognitive activation was usually recorded in the context of the TBD model as a generic quality dimension. Against this background, our study aimed to contribute to the investigation of the validity of student perceptions, considering subject-related characteristics of teaching quality and an opportunity-use model.

Table 2 Predicting student perceptions of their own learning and understanding and of teachers' comprehension orientation

	Model 1: Student: own learning process (Pyth)		Model 2: Student: own understanding (Pyth)		Model 3: Student: teacher's overall comprehension orientation (T4)	
	β	*SE*	β	*SE*	β	*SE*
Class-level variables						
Mean achievement (Pyth pretest)	−0.12*	(0.05)	−0.18**	(0.06)	–	
Mean achievement (math t1)	–		–		*nss*	
Expert ratings: Quality (Pyth)	0.16**	(0.06)	0.12*	(0.05)	*nss*	
Individual-level variables						
Achievement (Pyth pretest)	0.13**	(0.05)	0.16**	(0.05)	–	
Achievement (math t1)	–		–		*nss*	
Math-related interest	0.25**	(0.04)	0.27**	(0.04)	0.11**	(0.04)
General cognitive abilities	*nss*		0.15**	(0.04)	*nss*	

Note. β standardized HLM regression weight, *SE* standard error

** $p < .01$, * $p < 0.05$, *nss* not statistically significant

Table 3 Predicting students' achievement gains

	Model 4: Achievement: Pyth Post-test		Model 5: Achievement: Pyth posttest		Model 6: Achievement (end of school year)	
	β	*SE*	β	*SE*	β	*SE*
Class-level variables						
Mean achievement (Pyth pretest)	0.31**	(0.06)	0.33**	(0.07)	–	
Mean achievement (math t1)	–		–		*nss*	
Individual-level variables						
Achievement (Pyth pretest)	0.15**	(0.03)	0.15**	(0.03)	–	
Achievement (math t1)	–		–		0.22**	(0.05)
Math-related interest	0.08*	(0.03)	0.07**	(0.03)	0.13**	(0.03)
General cognitive abilities	0.22**	(0.03)	0.20**	(0.03)	0.26**	(0.04)
Student: own learning process (Pyth)	0.09**	(0.03)	–		–	
Student: own understanding (Pyth)	–		0.11**	(0.03)	–	
Student: teacher's overall comprehension orientation	–		–		0.06*	(0.03)

Note. β standardized HLM regression weight, *SE* standard error

** $p < .01$, * $p < 0.05$, *nss* not statistically significant

In summary, our results show a remarkable correlation between observer and student judgments at the *class level* when the observers' judgments refer to subject- specific quality features of understanding-oriented teaching and the students' judgments refer to the related learning processes and their understanding. Observer ratings are also reflected in *individual* student ratings of their own understanding and learning. Moreover, individual student ratings of their own learning processes and comprehension predict learning success when important learning prerequisites (interest, general cognitive ability, and prior knowledge) are being controlled.

6.1 Teaching quality as a co-production of teachers and students: the role of opportunity-use models in assessing subject-specific characteristics of instructional quality

6.1.1 Convergent validity of students' perceptions

According to opportunity-use models, the extent to which teaching promotes and supports students' subject-related understanding and learning processes does not result directly and necessarily from the quality of the learning opportunities in terms of teacher actions and tasks but rather from an interaction between these learning opportunities and the quality of how these are used. Therefore, from a theoretical viewpoint, teaching quality should be seen as a co-production of teachers and students (Cai et al., 2020; Fend, 1998; Reusser & Pauli, 2010; Vieluf & Klieme, 2023). From a methodological viewpoint, the question arises as to how the quality of use can be adequately measured, especially for characteristics that relate to students' understanding and learning processes. While observers can directly assess quality features of the offer, based on the features of teacher action, they have limited access to the students' mental processes and thus to the use of the offering. One promising possibility was explored by Prediger et al. (2023), who recorded active participation in class discussions as an indicator of usage. One problem could be that students may be cognitively active but not verbally involved in the interaction. Therefore, it makes sense to include the students' perspective as well (Vieluf, 2022).

Comparison of student and expert perceptions at the class level Previous studies have recorded student and observer assessments and, in some cases, compared them with each other, primarily for characteristics of the learning opportunities offered by the teacher. In contrast to these studies, which found rather low correlations between expert and student assessments of cognitive activation (Fauth et al., 2020), our experts focused on subject-specific learning opportunities and the students on their use. The substantial correlations between the two perspectives can be explained, on the one hand, by the fact that by focusing on "subject-specific quality", our experts specifically assessed the aspects of instruction that students need to build a robust understanding of the content (Pythagorean theorem), while the students assessed the related learning and understanding processes. It is undoubtedly easier for students to adequately assess their own learning and understanding of the subject matter than the subject-specific quality of instruction since they do not need subject didactic expertise to do so. The only requirement is that they can realistically assess their understanding and learning. On the other hand, it can be assumed

that the observers did not base their assessment of subject-specific quality exclusively on features of teacher behavior or task setting but also drew conclusions about students' processes of understanding. The quality assessments were related to clearly defined and content-standardized analysis units, which was not the case in most previous studies. This could also have contributed to a higher level of agreement between the expert and student assessments. Overall, the students' and observers' judgments in our study can be viewed as complementary sources for assessing the subject-specific quality of the instructional unit, understood as an interplay of offer and use.

Compared to the correlations between the student and observer perspectives related to the Pythagorean unit, the correlation is somewhat lower but still significant at the class level when students assess the general comprehension orientation of a mathematics teacher. Note that the student and observer judgments refer to different time periods and were collected at different times. While the observer judgments refer to the three Pythagorean lessons, the students retrospectively judged their mathematics teacher's general comprehension orientation at the end of the school year. As mentioned earlier, previous studies have mostly found little or no correlation between the perspectives on the cognitive aspects of learning support (see 2.2). In contrast, our observers did not assess generic but subject-specific quality characteristics.

Predicting individual student assessments: the role of student and class characteristics While individual students' perceptions of learning and understanding can be predicted by observer ratings of subject-specific instructional quality (Table 2), the best predictor of students' individual self-assessments is not the quality of teaching as assessed by the experts but the interest of the students. This finding is consistent with the research that has shown that students' perceptions of instruction are influenced by individual student characteristics (cf. Herbert et al., 2022; Wang & Eccles, 2016). This might also be the case for students' perceptions of their own cognitive processes and comprehension (Merk et al., 2021).

The finding that students' self-assessments of their learning and comprehension are also influenced by individual prior knowledge can be explained by learning psychology, which stresses the active, constructive, and cumulative character of learning processes (e.g. Aebli, 1983; Chi & Wylie, 2014; Reusser, 2006). The more and better interlinked the content-related prior knowledge is, the more successful the understanding and learning will be. The negative effect of mean prior knowledge at the class level is rather surprising. Two explanations are possible. First, it could be the result of the reference group effect (Marsh, 2005), as it can be assumed that comparisons with the class play a role when assessing one's own understanding and learning processes. The self-assessment may be more critical in comparison with a more capable group than in comparison with a less capable group, regardless of how well the learning content was understood and how successful one's own learning process was. A second possible explanation is that teachers adapt their teaching to the performance of the class and increase the expectations of the students depending on the level of the learning group. From this viewpoint, the negative effect could result from being confronted (in the more efficient classes) with higher expectations, more difficult tasks to solve, and more demanding teacher questions. Both explanations may also apply simultaneously.

6.1.2 Prognostic validity

Students' self-assessment of their learning and understanding also predicts their learning success. However, the effects of student perceptions are comparatively small. In contrast, the effects of the general cognitive ability as well as of prior knowledge on student and class level are, as expected, greater in terms of the short-term learning progress in the Pythagorean unit, which is equally positive at both levels.

Overall, previous research results have ascribed some prognostic validity to students' perceptions of instructional characteristics in predicting learning outcomes, with inconsistent findings for cognitive activation (cf. 2.2.2). In contrast to this research, the students in our case assessed their own learning processes and understanding and not the quality of teaching in terms of teacher behavior or the tasks. In this context, the question arises as to what extent learners can realistically rate their own learning processes and comprehension. Our results indicate that this is the case, at least to a certain degree, even if the relations are rather weak and the understanding was only recorded with a single item. This finding is consistent with previous research on self-perception of learning and comprehension processes (e.g. Lingel et al., 2019). For example, Nuthall and Alton-Lee (1990) showed, using data from student interviews, that even young students can describe their understanding and learning processes in a differentiated and precise manner.

6.2 Limitations, future research, and conclusions

Several features of our study require a discussion of its limitations. First, the analyses relate to a relatively small sample of classes and teachers that is not representative. Since participating in this study required considerable effort from the teachers and their willingness to have their lessons videotaped several times, it can be assumed that the participating teachers were a positive selection of particularly motivated

and perhaps particularly competent teachers. Therefore, the results are not generalizable to all teachers and students in lower secondary schools in the two participating countries. Second, the data were collected in the 2002/2003 school year, around 20 years ago, as part of the "Pythagoras study" (Klieme et al., 2009). This multi-faceted study generated a very rich database of test, survey, and video data, which still contains potential for further analysis from new perspectives. This can be achieved using a wide range of available video codes and ratings and by capitalizing on previous analyses. However, the age of the data raises the question of the validity of such analyses and their findings. The development of digital media over the last 20 years, among other things, has not only changed the world in which young people live, but has presumably also changed the design of mathematics lessons, for example the use of media. Nevertheless, we still consider our analyses and their results to be justifiable today. This is particularly because media use is a surface characteristic that says little about the quality of teaching, whereas our analyses focused on the deep structural quality of the lessons (cf. 2.1). If we were to analyze our lessons in terms of their design (e.g. methods, media) and compare them with today's lessons on the introduction of the Pythagorean theorem, we would probably find different patterns or frequencies and thus describe the design of lessons then and now differently. In contrast, the analyses presented here are not aimed at describing mathematics education (then or now) per se, but rather at investigating deep structure dimensions and relationships of teaching–learning quality (Klieme et al., 2009). The extent to which such relationships are influenced by changes in the design of lessons in terms of surface features such as methods and media is an open question that cannot be answered based on our analyses. In any case, it would be interesting and important to replicate our results based on current and representative data sets.

A third limitation is that the students' own perceived understanding was only recorded with a single item. It is therefore important that this variable is considered together with the other variable on the self-assessment of learning processes and understanding, which is based on a scale and with which it correlates, as expected (see Table A1, Appendix). It should also be noted that, in contrast to the students' perceptions of their own cognitive activity and their understanding, the students' perceptions of the teacher's comprehension orientation were only surveyed at the end of the school year.

Finally, since our analyses focused not only on subject-specific but also on content-specific characteristics of teaching quality, the results apply only to this content (i.e. the introduction of Pythagoras' theorem). It would be interesting to replicate our analyses with another rating instrument that measures the subject-specific characteristics of teaching quality. As mentioned in 2.2.1, numerous instruments have now been developed that cannot be listed here (overview e.g. in Praetorius & Charalambous, 2018; Schlesinger & Jentsch, 2016). Instruments based on the TBDs (e.g. the instrument developed for TEDS-Instruct, cf. Schlesinger et al., 2018) would be particularly interesting, as TBDs were also the basis for the Pythagoras study (Lipowsky et al., 2009) and are increasingly gaining international acceptance as a suitable framework for recording student assessments (cf. Herbert et al., 2022; Senden et al., 2023). One question of growing interest is whether and how the idea of EoU can also be transferred to other content in mathematics education (Korntreff & Prediger, 2022).

In the context of opportunity-use models of teaching and learning, it could also be interesting to have the students rate not only their use of the provided learning opportunities (in the present case: quality of the understanding and learning processes) but also the characteristics of the offer itself. However, due to the subject-specific knowledge discussed above, the question arises as to what extent such a rating would be possible and useful. In the present study, we refrained from doing so because we found assessments of students' own learning more promising as an indicator of their use of the learning opportunities than complementary judgments to the observer ratings. It was also important that the students could complete the questionnaire as spontaneously as possible and with as little time as possible at the end of the teaching unit.

In summary, our study contributes to a more nuanced picture of the potential of student perspective in assessing the deep structural quality of teaching. This is done by showing that the perspectives of students (by looking at the quality of their learning processes and understanding) and observers (by looking at the subject-didactic quality of the learning opportunities provided by the teachers) are two complementary sides of the same coin. Both are necessary to obtain a more complete picture of what is happening in the classroom.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11858-024-01561-3.

Funding Open access funding provided by University of Fribourg German Research Foundation (grant number: KL 1057/1-2 to Eckhard Klieme); Swiss National Science Foundation (grant number 1114-63564.00/1 to Kurt Reusser and Christine Pauli).

Declarations

Competing interests None.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are

included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Aebli, H. (1980/1981). *Denken, das Ordnen des Tuns [Thinking: The ordering of doing, Vol. I,II)*. Klett-Cotta.

Aebli, H. (1983). *Zwölf Grundformen des Lehrens [Twelve basic forms of teaching]*. Klett-Cotta.

Brunner, E. (2018). Qualität von Mathematikunterricht: Eine Frage der Perspektive. *Journal für Mathematik-Didaktik, 39*(2), 257–284. https://doi.org/10.1007/s13138-017-0122-z

Cai, J., Morris, A., Hohensee, C., Hwang, S., Robison, V., Cirillo, M., Kramer, S. L., Hiebert, J., & Bakker, A. (2020). Maximizing the quality of learning opportunities for every student. *Journal for Research in Mathematics Education, 51*(1), 12–25. https://doi.org/10.5951/jresematheduc.2019.0005

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*(3), 355–366. https://doi.org/10.1007/s11858-018-0914-8

Cheng, Q., Shen, J., & Zhang, S. (2023). Comparing perceived and observed instructional practices and their predictive power for student mathematics achievement: An analysis of Shanghai data from OECD global teaching inSights. *Asian Journal for Mathematics Education*, 27527263231210322. https://doi.org/10.1177/27527263231210322

Chi, M. T. H., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist, 49*(4), 219–243. https://doi.org/10.1080/00461520.2014.965823

Clausen, M. (2002). *Unterrichtsqualität: Eine Frage der Perspektive? Empirische Analysen zur Übereinstimmung, Konstrukt- und Kriteriumsvalidität [Instructional quality: A matter of perspective?]*. Waxmann.

De Jong, R., & Westerhof, K. J. (2001). The quality of student ratings of teacher behaviour. *Learning Environments Research, 4*(1), 51–85. https://doi.org/10.1023/A:1011402608575

Dreher, A., & Leuders, T. (2021). Fachspezifität von Unterrichtsqualität – aus der Perspektive der Mathematikdidaktik [Subject-specificity of instructional quality—From the perspective of mathematics education]. *Unterrichtswissenschaft, 49*(2), 285–292. https://doi.org/10.1007/s42010-021-00116-9

Drollinger-Vetter, B. (2011). *Verstehenselemente und strukturelle Klarheit. Fachdidaktische Qualität der Anleitung von mathematischen Verstehensprozessen im Unterricht [Elements of comprehension and structural clarity]*. Waxmann.

Drollinger-Vetter, B., & Lipowsky, F. (2006). Fachdidaktische Qualität der Theoriephasen [Quality of theory phases in mathematics lessons]. In I. Hugener, C. Pauli, & K. Reusser (Eds.), *Videoanalysen (= Teil 3 der Dokumentation Erhebungs- und Auswertungsinstrumente zur schweizerisch-deutschen Videostudie "Unterrichtsqualität, Lernverhalten und mathematisches Verständnis", hrsg. von E. Klieme, C. Pauli & K. Reusser)* (pp. 189–205). GFPF/DIPF. https://doi.org/10.25656/01:3130

Drollinger-Vetter, B., Lipowsky, F., Pauli, C., Reusser, K., & Klieme, E. (2006). Cognitive level in problem segments and theory segments. *ZDM–Mathematics Education*, *38*(5), 399–412.

Fauth, B., Göllner, R., Lenske, G., Praetorius, A.-K., & Wagner, W. (2020). Who sees what? Conceptual considerations on the measurement of teaching quality from different perspectives. *Zeitschrift für Pädagogik*, *66. Beiheft*, 255–268.

Fauth, B., Decristan, J., Rieser, S., Klieme, E., & Büttner, G. (2014). Student ratings of teaching quality in primary school: Dimensions and prediction of student outcomes. *Learning and Instruction*, *29*(0), 1–9. https://doi.org/10.1016/j.learninstruc.2013.07.001

Fend, H. (1998). *Qualität im Bildungswesen. Schulforschung zu Systembedingungen, Schulprofilen und Lehrerleistung [Quality in Education. Research on the quality of school system, school profiles, and teacher performance]*. Juventa.

Göllner, R., Fauth, B., & Wagner, W. (2021). Student ratings of teaching quality dimensions: Empirical findings and future directions. In W. Rollett, H. Bijlsma, & S. Röhl (Eds.), *Student feedback on teaching in schools: Using student perceptions for the development of teaching and teachers* (pp. 111–122). Springer International Publishing. https://doi.org/10.1007/978-3-030-75150-0_7

Gravemeijer, K., Stephan, M., Julie, C., Lin, F.-L., & Ohtani, M. (2017). What mathematics education may prepare students for the society of the future? *International Journal of Science and Mathematics Education, 15*(1), 105–123. https://doi.org/10.1007/s10763-017-9814-6

Heller, K. A., & Perleth, C. (2000). *Kognitiver Fähigkeitstest für 4. bis 12. Klassen, Revision (KFT 4–12+R) [Cognitive ability test]*. Beltz Test.

Helmke, A. (2003). *Unterrichtsqualität - erfassen, bewerten, verbessern [Assessing, evaluating, and enhancing instructional quality]*. Kallmeyer.

Herbert, B., Fischer, J., & Klieme, E. (2022). How valid are student perceptions of teaching quality across education systems? *Learning and Instruction*, *82*, 101652. https://doi.org/10.1016/j.learninstruc.2022.101652

Hiebert, J., & Carpenter, T. P. (1992). Learning and teaching with understanding. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 65–97). Macmillan.

Hiebert, J., & Grouws, D. A. (2007). The effects of classroom mathematics teaching on students' learning. In F. K. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 371–404). Information Age Publishing.

Hugener, I., Pauli, C., Reusser, K., Lipowsky, F., Rakoczy, K., & Klieme, E. (2009). Teaching patterns and learning quality in Swiss and German mathematics lessons. *Learning and Instruction*, *19*(1), 66–78. https://doi.org/10.1016/j.learninstruc.2008.02.001

Jansen, N. C., Decristan, J., & Fauth, B. (2022). Individuelle Nutzung unterrichtlicher Angebote – Zur Bedeutung von Lernvoraussetzungen und Unterrichtsbeteiligung [Individual use of instruction–On the relevance of students' characteristics and participation in classroom discourse]. *Unterrichtswissenschaft, 50*(2), 157–183. https://doi.org/10.1007/s42010-021-00141-8

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

Koedinger, K. R., Corbett, A. T., & Perfetti, C. (2012). The Knowledge-Learning-Instruction (KLI) framework: Bridging the science-practice chasm to enhance robust student learning. *Cognitive Science, 36*(5), 757–798. https://doi.org/10.1111/j.1551-6709.2012.01245.x

Korntreff, S., & Prediger, S. (2022). Verstehensangebote von YouTube-Erklärvideos – Konzeptualisierung und Analyse am Beispiel algebraischer Konzepte [Conceptual Learning Opportunities of Instructional YouTube Videos – Conceptualization and Analysis for the Case of Algebraic Concepts]. *Journal Für*

Mathematik-Didaktik, 43(2), 281–310. https://doi.org/10.1007/s13138-021-00190-7

Kunter, M., & Baumert, J. (2006). Who is the expert? Construct and criteria validity of student and teacher ratings of instruction. *Learning Environments Research, 9*(3), 231–251. https://doi.org/10.1007/s10984-006-9015-7

Kunter, M., & Voss, A. (2013). The model of instructional quality in COACTIV: A multicriteria analysis. In M. Kunter, J. Baumert, D. Blum, U. Klusmann, S. Krauss, & M. Neubrand (Eds.), *Cognitive activation in the mathematics classroom and professional competence of teachers* (pp. 97–124). Springer.

Lenske, G., & Praetorius, A.-K. (2020). Schülerfeedback - was steckt hinter dem Kreuz auf dem Fragebogen? [Student feedback–what is the basis for choosing an answer on a questionnaire?] *Empirische Pädagogik, 34*(1), 11–29.

Lindmeier, A., & Heinze, A. (2020). Die fachdidaktische Perspektive in der Unterrichtsqualitätsforschung: (bisher) ignoriert, implizit enthalten oder nicht relevant? [The subject-specific perspective in teaching quality research: (so far) ignored, implicitly included or not relevant?]. *Zeitschrift für Pädagogik, 66. Beiheft*, 255–268.

Lingel, K., Lenhart, J., & Schneider, W. (2019). Metacognition in mathematics: do different metacognitive monitoring measures make a difference? *ZDM–Mathematics Education*, *51*(4), 587–600. https://doi.org/10.1007/s11858-019-01062-8

Lipowsky, F., Drollinger-Vetter, B., Hartig, J., & Klieme, E. (2006). *Dokumentation der Erhebungs- und Auswertungsinstrumente zur schweizerisch-deutschen Videostudie "Unterrichtsqualität, Lernverhalten und mathematisches Verständnis". 2. Leistungstests [Achievement tests]*. GFPF. https://doi.org/10.25656/01:3107

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

Lüdtke, O., Robitzsch, A., Trautwein, U., & Kunter, M. (2009). Assessing the impact of learning environments: How to use student ratings of classroom or school characteristics in multilevel modeling. *Contemporary Educational Psychology, 34*(2), 120–131. https://doi.org/10.1016/j.cedpsych.2008.12.001

Marsh, H. W. (2005). Big-fish-little-pond effect on academic self-concept. *Zeitschrift Für Pädagogische Psychologie, 19*(3), 119–129. https://doi.org/10.1024/1010-0652.19.3.119

Mayer, R. E. (2009). Constructivism as a theory of learning versus constructivism as a prescription for instruction. In S. Tobias & T. Duffy (Eds.), *Constructivist instruction: Success or failure?* (pp. 184–200). Routledge.

Merk, S., Batzel-Kremer, A., Bohl, T., Kleinknecht, M., & Leuders, T. (2021). Nutzung und Wirkung eines kognitiv aktivierenden Unterrichts bei nicht-gymnasialen Schülerinnen und Schülern [Use and effects of cognitively activating instruction on non-high school students]. *Unterrichtswissenschaft, 49*(3), 467–487. https://doi.org/10.1007/s42010-021-00101-2

Nuthall, G., & Alton-Lee, A. (1990). Research on teaching and learning: Thirty years of change. *The Elementary School Journal, 90*(5), 547–570. https://doi.org/10.1086/461632

Patrick, H., Mantzicopoulos, P., & Sears, D. (2012). Effective classrooms. In K. R. Harris, S. Graham, & T. Urdan (Eds.), *Educational Psychology Handbook* (Vol. 2, pp. 443–470). American Psychological Association. https://doi.org/10.1037/13274-018

Praetorius, A.-K., & Charalambous, C. Y. (2018). Classroom observation frameworks for studying instructional quality: looking back and looking forward. *ZDM–Mathematics Education*, *50*(3), 535–553. https://doi.org/10.1007/s11858-018-0946-0

Praetorius, A.-K., Klieme, E., Herbert, B., & Pinger, P. (2018). Generic dimensions of teaching quality: the German framework of Three Basic Dimensions. *ZDM–Mathematics Education*, *50*(3), 407–426. https://doi.org/10.1007/s11858-018-0918-4

Prediger, S., Götze, D., Holzäpfel, L., Rösken-Winter, B., & Selter, C. (2022). Five principles for high-quality mathematics teaching: Combining normative, epistemological, empirical, and pragmatic perspectives for specifying the content of professional development. *Frontiers in Education*, *7*. https://doi.org/10.3389/feduc.2022.969212

Prediger, S., Erath, K., Quabeck, K., & Stahnke, R. (2023). Effects of interaction qualities beyond task quality: Disentangling instructional support and cognitive demands. *International Journal of Science and Mathematics Education*. https://doi.org/10.1007/s10763-023-10389-4

Rakoczy, K., Buff, A., & Lipowsky, F. (2005). *Dokumentation der Erhebungs- und Auswertungsinstrumente zur schweizerisch-deutschen Videostudie "Unterrichtsqualität, Lernverhalten und mathematisches Verständnis". 1. Befragungsinstrumente [Questionnaires]*. GFPF. https://doi.org/10.25656/01:3106

Rakoczy, K., Klieme, E., Drollinger-Vetter, B., Lipowsky, F., Pauli, C., & Reusser, K. (2007). Structure as a quality feature in mathematics instruction: Cognitive and motivational effects of a structured organisation of the learning environment vs. a structured presentation of learning content. In M. Prenzel (Ed.), *Studies on the educational quality of schools. The final report on the DFG Priority Programme* (pp. 101–120). Waxmann.

Raudenbush, S. W., Rowan, B., & Cheong, Y. F. (1993). Higher order instructional goals in secondary school: Class, teacher and school influences. *American Educational Research Journal, 30*(3), 523–553. https://doi.org/10.3102/00028312030003523

Reusser, K. (2005). Problemorientiertes Lernen - Tiefenstruktur, Gestaltungsformen, Wirkung [Problem-based learning–deep structure, forms of design, effects]. *Beiträge zur Lehrerbildung*, *23*(2), 159–182. https://doi.org/10.25656/01:13570

Reusser, K. (2006). Konstruktivismus - vom epistemologischen Leitbegriff zur Erneuerung der didaktischen Kultur [Constructivism: from a key epistemological concept to a new instructional culture]. In M. Baer, M. Fuchs, P. Füglister, K. Reusser, & H. Wyss (Eds.), *Didaktik auf psychologischer Grundlage. Von Hans Aeblis kognitionspsychologischer Didaktik zur modernen Lehr- und Lernforschung* (pp. 151–168). hep.

Reusser, K., & Reusser-Weyeneth, M. (1997). Verstehen als psychologischer Prozess und als didaktische Aufgabe [Understanding: psychological process and didactic task]. In K. Reusser & M. Reusser-Weyeneth (Eds.), *Verstehen. Psychologischer Prozess und didaktische Aufgabe* (2 ed., pp. 9–35). Huber.

Reusser, K., & Pauli, C. (2010). Unterrichtsgestaltung und Unterrichtsqualität - Ergebnisse einer internationalen und schweizerischen Videostudie zum Mathematikunterricht: Einleitung und Überblick [Lesson design and quality of instruction: Results of an international and Swiss video study on mathematics teaching. Introduction and overview]. In K. Reusser, C. Pauli, & M. Waldis (Eds.), *Unterrichtsgestaltung und Unterrichtsqualität – Ergebnisse einer internationalen und schweizerischen Videostudie zum Mathematikunterricht* (pp. 9–32). Waxmann.

Rieser, S., & Decristan, J. (2023). Kognitive Aktivierung in Befragungen von Schülerinnen und Schülern [Cognitive activation in student questionnaires – Distinguishing between the potential for cognitive activation and individual cognitive activation]. *Zeitschrift Für Pädagogische Psychologie*. https://doi.org/10.1024/1010-0652/a000359

Scherer, R., & Gustafsson, J.-E. (2015). Student assessment of teaching as a source of Information about aspects of teaching quality in multiple subject domains: An application of multilevel bifactor structural equation modeling. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.01550

Scherer, R., Nilsen, T., & Jansen, M. (2016). Evaluating individual students' perceptions of instructional quality: An investigation of their factor structure, measurement invariance, and relations to educational outcomes. *Frontiers in Psychology*, *7*(110). https://doi.org/10.3389/fpsyg.2016.00110

Schlesinger, L., & Jentsch, A. (2016). Theoretical and methodological challenges in measuring instructional quality in mathematics education using classroom observations. *ZDM–Mathematics Education*, *48*(1), 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*(3), 475–490. https://doi.org/10.1007/s11858-018-0917-5

Schoenfeld, A. H. (2018). Video analyses for research and professional development: the teaching for robust understanding (TRU) framework. *ZDM–Mathematics Education*, *50*(3), 491–506. https://doi.org/10.1007/s11858-017-0908-y

Senden, B., Nilsen, T., & Teig, N. (2023). The validity of student ratings of teaching quality: Factorial structure, comparability, and the relation to achievement. *Studies in Educational Evaluation*, *78*, 101274. https://doi.org/10.1016/j.stueduc.2023.101274

Vieluf, S. (2022). Wie, wann und warum nutzen Schüler*innen Lerngelegenheiten im Unterricht? Eine übergreifende Diskussion der Beiträge zum Thementeil [How, when and why do students use learning opportunities in the classroom? An overarching discussion of the contributions to the topical focus section]. *Unterrichtswissenschaft, 50*(2), 265–286. https://doi.org/10.1007/s42010-022-00144-z

Vieluf, S., & Klieme, E. (2023). Teaching Effectiveness Revisited Through the Lens of Practice Theories. In A.-K. Praetorius & C. Y. Charalambous (Eds.), *Theorizing Teaching: Current Status and Open Issues* (pp. 57–95). Springer International Publishing. https://doi.org/10.1007/978-3-031-25613-4_3

Wagner, W., Göllner, R., Helmke, A., Trautwein, U., & Lüdtke, O. (2013). Construct validity of student perceptions of instructional quality is high, but not perfect: Dimensionality and generalizability of domain-independent assessments. *Learning and Instruction*, *28*(0), 1–11. https://doi.org/10.1016/j.learninstruc.2013.03.003

Waldis, M., Grob, U., Pauli, C., & Reusser, K. (2010). Der Einfluss der Unterrichtsgestaltung auf Fachinteresse und Mathematikleistung [The influence of instruction on interest and mathematics achievement]. In K. Reusser, C. Pauli, & M. Waldis (Eds.), *Unterrichtsgestaltung und Unterrichtsqualität – Ergebnisse einer internationalen und schweizerischen Videostudie zum Mathematikunterricht* (pp. 209–251). Waxmann.

Wallace, T. L., Kelcey, B., & Ruzek, E. (2016). What can student perception surveys tell us about teaching? Empirically testing the underlying structure of the Tripod student perception survey. *American Educational Research Journal, 53*(6), 1834–1868. https://doi.org/10.3102/0002831216671864

Wang, M.-T., & Eccles, J. S. (2016). Multilevel predictors of math classroom climate: A comparison study of student and teacher perceptions. *Journal of Research on Adolescence, 26*(3), 617–634. https://doi.org/10.1111/jora.12153

Wertheimer, M. (1945). *Productive thinking*. Harper.

Wisniewski, B., Zierer, K., Dresel, M., & Daumiller, M. (2020). Obtaining secondary students' perceptions of instructional quality: Two-level structure and measurement invariance. *Learning and Instruction, 66*, 101303. https://doi.org/10.1016/j.learninstruc.2020.101303

Wu, M. L., Adams, R. J., & Wilson, M. R. (1997). *ConQuest: Multi-aspect test software [computer program]*. Australian Council for Educational Research.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.