ZDM – Mathematics Education
https://doi.org/10.1007/s11858-024-01552-4

ORIGINAL PAPER

Differential instructional qualities despite equal tasks: Relevance of school contexts for subdomains of cognitive demands

Kim Quabeck[1] · **Kirstin Erath**[1,3] · **Susanne Prediger**[1,2]

Accepted: 19 January 2024

Abstract
Cognitive demand is a crucial dimension of instructional quality. Its heterogenous operationalizations call for refined investigations, with respect to discursive richness (generic conceptualizations) and conceptual richness (subject-related conceptualizations). Considering not only teachers' intended cognitive activation (operationalized, e.g., by tasks), but also the enacted activation and individual students' participation as realized in the interaction, raises the question of how far the interaction quality is associated with students' prerequisites, school context, and class composition. In this paper, we present a video study of leader-led small-group instruction (in 49 groups of 3–6 middle school students each) with the same fraction tasks, so that differences in interaction quality can be scrutinized in generic and subject-related conceptualizations. In spite of equal task quality, large differences occurred in interaction quality across heterogenous class compositions. The regression analyses revealed that the enacted activation and individual participation were significantly associated with the school context (of higher-tracked and lower-tracked schools), but much less with individual learning prerequisites. These findings reveal the need to capture students' collective and individual engagement in cognitive demands in the interaction and in generic and subject-related conceptualizations and to systematically investigate their association with class composition.

Keywords Subject-related instructional quality · Interaction quality · Differential learning opportunities · Operationalizations

1 Introduction: disentangling instructional quality in different school contexts

Instructional quality dimensions of cognitive demands and instructional support have been shown to have a significant impact on students' learning gains (Bostic et al., 2021; Pianta & Hamre, 2009; Praetorius et al., 2018). As research surveys have shown (Bostic et al., 2021; Mu et al., 2022), most existing quality coding protocols have used comprehensive ratings that holistically combine various subdomains of *task quality* and *interaction quality*, with heterogeneous conceptualizations and operationalizations. Researchers have therefore called for disentangling these conceptualizations in three directions:

- Researching the role of subject-related and generic measures of instructional quality (Brunner, 2018; Praetorius & Charalambous, 2018; Schlesinger et al., 2018)
- Distinguishing supply and use (Brühwiler & Blatchford, 2011) into teachers' intended activation (as given, e.g., by tasks or teacher moves), enacted activation (i.e., the class engagement after tasks and moves), and individual participation (of each individual student; Ing & Webb, 2012; Sedova et al., 2019)
- Making explicit and distinguishing different operationalizations (Ing & Webb, 2012; Mu et al., 2022; Praetorius & Charalambous, 2018).

With our project, we have been contributing to this *research agenda of disentangling* conceptualizations by investigating the interaction quality (enacted activation and individual participation) in classrooms in which the task quality (intended activation) and planned teacher moves

✉ Susanne Prediger
prediger@dzlm.de

[1] TU Dortmund University, Dortmund, Germany

[2] IPN Leibniz Institute for Science and Mathematics Education, Berlin, Germany

[3] Martin Luther University of Halle-Wittenberg, Halle, Germany

Published online: 04 April 2024

were held constant. We conceptualize and operationalize generic subdomains of cognitive demands and instructional support as well as subject-related subdomains to compare their prevalence in classrooms (Quabeck et al., 2023; Prediger et al., 2023).

As requested by Fauth et al. (2021), we have continued the investigation of in-depth differences of instructional quality (operationalized by different quality features of enacted activation and individual participation) to study how their prevalence is associated with class composition, school context, and students' heterogeneous prerequisites. Differential instructional qualities for schools and students with heterogeneous backgrounds have often been shown to create unequal learning opportunities and unequal learning outcomes (DIME, 2007). So far, however, we have not been able to identify studies that have analyzed differential (generic and subject-related) qualities of *video-recorded interaction* in classes with the same selection of tasks. By keeping tasks identical across 49 teacher-led middle school small-groups, a focus on the influence of students' heterogeneous prerequisites and differential school contexts on interaction quality becomes possible. Therefore, we pursue the following research question (to be refined later):

To what extent can differential generic and subject-related quality features be predicted by students' heterogenous prerequisites and school contexts?

In Sect. 2, we briefly report on heterogeneous conceptualizations and operationalizations of cognitive demands in existing quality coding protocols and present our coding protocol for disentangling them. We embed our research question in existing research on differential learning milieus and class composition effects and the supply-use model (Brühwiler & Blatchford, 2011; Helmke, 2009). In Sect. 3, we present the methods of data gathering and data analysis in the developed coding protocol. In Sect. 4, we show that, indeed, school contexts reveal substantial differences in interaction qualities even when task quality is held constant and show that these differences are not explained solely by students' prerequisites. The findings are discussed in Sect. 5.

2 Theoretical background: conceptualizing and operationalizing interaction quality for studying differential learning milieus beyond task quality

2.1 Coding protocols for instructional quality covering task quality and interaction quality

In many international studies, empirical evidence has been provided that students' learning gains significantly depend on the quality of instruction (Brophy, 2000; Cai et al., 2020; Bostic et al., 2021). This applies to more *generic coding protocols* (e.g., Praetorius et al., 2018; Pianta & Hamre, 2009) and to more *subject-related coding protocols* (e.g., Blömeke et al., 2022; Hill et al., 2008).

In particular, the quality dimensions of *cognitive demands* and *instructional support* have strong effects on students' learning gains. As they overlap and have been used with multiple conceptualizations and operationalizations (Mu et al., 2022; Praetorius & Charalambous, 2018; Spreitzer et al., 2022), it has been emphasized that "we are only beginning to understand what makes a difference in terms of quality teaching" (OECD, 2020, p. 14). So, researchers called for further striving for depth in the ways instructional quality is measured and for comparing different school contexts.

With respect to cognitive demands, the insightful survey by Mu et al. (2022) revealed the heterogeneity of conceptualizations and operationalizations and calls for more conceptual clarity, distinguishing the cognitive demand of tasks from teachers' cognitively demanding facilitation and students' cognitive engagement. Whereas *task quality* has often been captured by subject-related quality features (e.g., Ni et al., 2018; Neubrand et al., 2013), *interaction quality* in teachers' facilitation and students' engagement has more often been captured by generic quality features (Bostic et al., 2021). For instructional support, the survey results suggest distinguishing the generic support of social relatedness from more subject-related support of competency and autonomy experience (Mu et al., 2022), both situated in interaction rather than task qualities.

Furthermore, quality dimensions in different coding protocols differ in their focus on teachers' *intended activation* (captured by the tasks and teacher moves, e.g., in Neubrand et al., 2013; Schlesinger et al., 2018), teachers' *enacted activation* (captured by some measure of class engagement, e.g., Lipowsky et al., 2009), or each *student's individual engagement* (captured mainly by individual talk time, e.g., Sedova et al., 2019). When class engagement or individual students' engagement is rated by some indicator of richness, this indicator can be operationalized by the richness of the *task* in which students engage (*task-based operationalizations*, e.g., Hill et al., 2008), the richness of a teacher move after which students engage (*move-based operationalizations,* e.g., Schoenfeld, 2018), or the richness of the interactively established discourse practices (*practice-based operationalizations*, e.g., Hill et al., 2008; Ing & Webb, 2012; OECD, 2020). Bostic et al. (2021) and Quabeck et al. (2023) provided an overview of a large heterogeneity of conceptualizations and operationalizations, for which Ing and Webb (2012) had already shown that these differences can lead to different quality judgments, so they should be transparently accounted for.

2.2 Disentangling conceptualizations and operationalizations of interaction quality

While different task qualities have been studied in depth with transparent subject-related conceptualizations and operationalizations, it is the *interaction quality* that is in major need of further disentanglement (Mu et al., 2022; Spreitzer et al., 2022).

With respect to overlapping and heterogenous conceptualizations and operationalizations of cognitive demand and instructional support, it is important to disentangle how different coding protocols conceptualized interactional quality. First, the existing coding protocols can be distinguished in capturing *intended activation*, *enacted activation*, and *individual students' participation*. Activation and participation can be simply measured by the talk time of teachers and students (early work of Flanders, 1970), but talk is only a necessary not a sufficient condition for student learning.

That is why a second distinction must be made according to the different conceptualizations of richness that have been identified as distinct and relevant in qualitative case studies on interaction (Lampert & Cobb, 2003; Walshaw & Anthony, 2008): (a) discursive richness, (b) conceptual richness, and (c) lexical richness. All three have been used in existing coding protocols, as the following brief summary reveals. Third, the existing coding protocols differ in their ways of operationalizing the measures.

Discursive richness has been captured in many subject-independent parts of coding protocols, but has been criticized as too simplifying (Pauli & Reusser, 2015) when teachers' intended activation is captured by move-based measures instead of enacted classroom practices (e.g., Stigler et al., 1999). Other studies have utilized a combination of moves and practices (e.g., Boston, 2012; Hill et al., 2008; Bostic et al., 2021) or have mainly focused on practices (e.g., students' and teachers' explanations; OECD, 2020) for capturing the enacted discursive richness in the interaction. Individual participation has been assessed by counting the number of students' utterances with reasoning (e.g., Sedova et al., 2019) or by calculating the length of student contribution in word- or time-related measurements (e.g., Lipowsky et al., 2009). These *task-based*, *move-based*, or *practice-based operationalizations* have still varied in terms of their rating or coding: They have been either rated roughly in time segments of different sizes (Praetorius & Charalambous, 2018) or by sentence-related (e.g., number of words spoken, Stigler et al., 1999) or time-related frequencies (e.g., Sedova et al., 2019).

The same applies to measures of *conceptual richness*, the most relevant subject-specific domain by which the quality of the knowledge negotiated has been captured, for instance, by rating the quality of the task implementation by teaching practice (Boston, 2012), by the enacted move demand (Pauli & Reusser, 2015), or in the conceptual practices that are co-constructed in the interaction (e.g., in some 4-point scale ratings by Hill et al., 2008). Even the conceptual practices co-constructed in the interaction have been operationalized diversely and have differed (Quabeck et al., 2023), for example, in the extent to which they mainly comprise teachers' enactment of mathematical richness in interaction (e.g., depths of the mathematics offered) or focus more on class engagement (e.g., level of student work). However, published coding manuals have scarcely explicated the exact operationalization bases (task, moves, or practices) for the quality assessment. Often, several bases have been mentioned and combined, yet the exact relation between combined bases has not been transparent.

By *lexical richness*, we refer to an aspect of instructional support that is particularly relevant for students from underprivileged backgrounds who still accomplish their academic language proficiency (Gibbons, 2002). Teachers' rich lexical support for students' vocabulary acquisition has been identified as productive in mathematics classrooms when embedded in rich discourse practices such as explaining meanings or arguing (Gibbons, 2002; Moschkovich, 2015). In our analytic framework to be presented in Section 3.5, we consider only vocabulary relevant for students' conceptual learning (Moschkovich, 2015), so this is a subject-related quality domain.

In this paper, we will show that these different kinds of richness are established differentially in different learning milieus. To accomplish this, the next subsection outlines the argument for why differential learning milieus are crucial to consider.

2.3 Differential learning milieus in different school contexts

Cai et al. (2020) emphasized the research need of "defining and measuring learning opportunities precisely enough" and elaborated that the "urgency of extending and refining the research on learning opportunities comes, in part, from the fact that high-quality learning opportunities are unequally distributed" (p. 13) in relation to various social background factors. Indeed, many large-scale assessments have documented the existence of unequal learning outcomes and have traced them back to *differential learning milieus* in educational systems with between-school tracking (Maaz et al., 2008) or differential learning opportunities with more subtle differences (DIME, 2007). In the international discourse about these findings, differential learning gains of different school contexts have been traced back to three bundles of effects: prerequisite effects, class composition effects, and institutional effects (Maaz et al., 2008).

Studies in which differential learning milieus have been explained by *prerequisite effects* have shown that the

differential learning gains can be statistically predicted by students' individual background factors such as immigrant status of the families, socioeconomic status, language proficiency, and prior mathematical knowledge (Becker et al., 2022). But this statistical perspective has not considered how the prerequisites interplay with the learning opportunities provided (DIME, 2007; Howe & Abedin, 2013). *Institutional effects* have been identified in educational systems with in-school or between-school tracking through differential curricula with less ambitious learning goals, less rich tasks, less qualified teachers, and/or different classroom cultures (Oaks et al., 1992; Maaz et al., 2008). Additionally, *class composition effects* have explained variances in learning gains within the same school types according to the percentages of students of underprivileged backgrounds (Becker et al., 2022).

The German educational system (in the federal state North-Rhine-Westphalia in which the current study is situated) has been shaped by procedures of early between-school tracking at the age of 10 years, with about 40% of students in higher-tracked schools and 60% in lower-tracked schools (and some variations in the other federal states), and strong social disparities in school attendance in lower and higher tracks. Even when students' individual background factors are controlled, German higher-tracked schools have tended to lead their students to higher learning gains than lower-tracked schools (across all states) have. Within the lower-tracked schools in North-Rhine-Westphalia, students who struggle in mathematics often experience additional ability grouping, so the at-risk school context within this educational system has tended to consist of those students for whom lower-tracked schools have failed to develop their mathematics understanding. The differences between these German school types have also influenced studies on instructional quality: For example, Blömeke et al. (2022) reported that some influences of quality features disappeared when controlling for school types, given the strong differences in classroom cultures and tasks. Lipowsky et al. (2007) found significantly higher conceptual and discursive richness of classroom talk in higher-tracked schools, and Baumert et al. (2010) found higher cognitive demands on the task level. Fauth et al. (2021) investigated effects of class composition on instructional quality in primary school science classrooms and found strong correlations of cognitive class composition for classroom management, but not for cognitive demand. These heterogeneous findings (found in different states) call for further analyses of the compositional effects on emerging learning opportunities, with more in-depth operationalizations of instructional quality, not only for task quality, but also for interaction quality.

One conjecture emerging from the comparison of study designs has been that the missing differences in cognitive demands in the study by Fauth et al. (2021) might be traced back to a constant task quality provided in the study, where all teachers shared the same primary science tasks. In contrast, the other studies might have mainly captured institutional effects of differential curricula and tasks that also influenced instructional quality.

Our study is designed to contribute to the analysis of differential instructional qualities for different school contexts and class compositions by keeping the task quality constant and by thoroughly analyzing the enacted activation and individual students' participation.

3 Methodological framework for analyzing differential learning gains

3.1 Overall research design

Given the potential relevance of different school contexts (Fauth et al. 2021; Pauli & Reusser, 2015), we refine the initial research question by starting first with a descriptive question RQ1 before studying the regression models in RQ2 that help to unfold whether prerequisite effects or class composition effects apply:

RQ1 How does the interaction in classes from different school contexts differ with respect to different generic and subject-related quality features?

RQ2 To what extent can differential generic and subject-related quality features of interaction be predicted by students' prerequisites and school contexts?

We situate our research within the methodological framework of the supply-use model (Brühwiler & Blatchford, 2011; Helmke, 2009), in which instruction is investigated with respect to teachers' supply and individual students' use (Prediger et al., 2023). Given that interaction is co-constructed by teachers and students (Lampert & Cobb, 2003; Walshaw & Anthony, 2008), we adapted the supply-use model in Fig. 1 by splitting teachers' supply into the task quality as *intended teacher activation* (being the same in all small groups, so not to be further analyzed here) and interactional supply as *teachers' enacted activation*. A student's individual use refers to *individual participation*, so we conceptualize eight subdomains of quality. We then study how their operationalizations into 14 quality features are predictable by students' individual prerequisites or the school contexts. All components are further explained in the next subsections (following the Method section in Prediger et al., 2023).

3.2 Constant task quality in a small-group fraction intervention

The data corpus originates from the intervention study MuM-Mesut, with a discursively, conceptually, and lexically

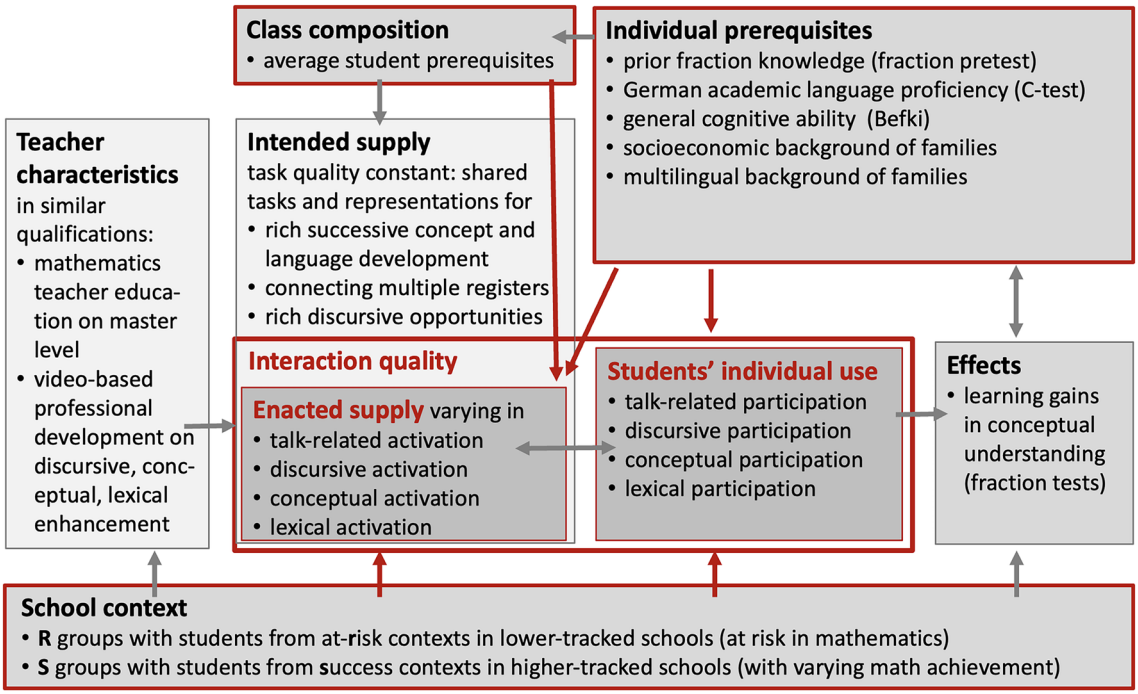

Fig. 1 Research design illustrated in the framework of an adapted supply-use model (adapted from Prediger et al., 2023, based on Helmke, 2009, in red the parts investigated in this paper)

rich intervention aiming at developing conceptual understanding of fractions and their operations and at developing a bridging language for explaining meanings (Prediger et al., 2022). The curriculum materials and teacher preparation focused on establishing conceptual richness using tasks and prepared teacher moves along a carefully designed conceptual learning trajectory and by connecting multiple representations. To establish discursive richness, tasks systematically invite students to engage in rich discourse practices and support their lexical development with language-responsive design principles (Moschkovich, 2015). Empirical evidence for the overall efficacy of the intervention was provided in a cluster-randomized controlled trial in which students in the intervention groups (n = 394) showed significantly higher learning gains than the control group (n = 195) with business-as-usual teaching (Prediger et al., 2022).

In order to capture each individual student's participation in detail, the instruction was organized in separate teacher-led small group instructions (3–6 students with one teacher) and spanned over five video-recorded sessions of 90 minutes each, taught by master and PhD students with mathematics teaching certificates. The video-based teacher preparation and weekly supervision meetings and the written manual detailed the task goals, typical student challenges and relevant questions and moves. The implementation check revealed strong fidelity in task sequences and overall time spent for the tasks, but with large differences in subtasks and prompts. By keeping the tasks (and representations and suggested teacher moves) nearly identical (and teachers' intended activation agreed upon in the teacher preparation), differences in teachers' enacted activation in the interaction and individual students' participation become observable in a fine-grained way.

3.3 Measures for individual prerequisites

The following quantitative measures were administered prior to the video-recorded interventions (Prediger et al., 2022):

- *Prior fraction knowledge.* Students' conceptual understanding of fractions (the dependent variable) was measured by a standardized fraction pretest, covering aspects of conceptual understanding and procedures with fractions (internal consistency of Cronbach's α = .82 with 25 items).
- *Academic language proficiency.* Students' academic language proficiency in the German language of instruction was measured by a C-Test, a widely used, economical, and valid measure with cloze texts to assess vocabulary and grammar knowledge of the language in complex situated ways (internal consistency of Cronbach's α = .788).
- *General cognitive ability.* Fluid intelligence was measured using a matrix test (BEFKI 7, with internal consistency of Cronbach's α = .78 in 16 items).
- *Multilingual background.* Multilingual students were those who reported speaking multiple family languages

Table 1 Descriptive data for the full sample and the subsamples in view

Variable M (SD) or percent	Sample of initial intervention R'+S' ($N = 589$)	Sample of video study Full video sample ($n = 210$)	…in at-risk school contexts R ($n = 83$)	… in successful school contexts S ($n = 127$)
Fraction pretest score	7.37 (3.78)	7.64 (3.8)	8.88 (3.17)	6.81 (3.97)
General cognitive ability	8.47 (3.32)	8.82 (3.8)	8.27 (2.74)	9.19 (4.34)
Academic language proficiency	38.26 (9.36)	40.04 (9.16)	37.19 (8.53)	41.94 (9.11)
Age	11.86 (1.18)	11.53 (1.15)	12.79 (0.62)	10.71 (0.49)
Multilingual background	57%	49%	52%	48%
SES: low/medium/high	26%/31%/43%	21%/32%/47%	36%/36%/28%	11%/30%/60%

(or a family language other than the language of instruction) in the survey.

- *Socioeconomic status.* Students' SES was measured using the book-at-home index levels, asking students how many books they have at home with example photos (re-test reliability of $r = 0.80$, level 1–5).

3.4 Sample and sampling

To compare different school milieus in our federal state, North-Rhine-Westphalia, we needed to ensure that we considered both, between-school tracking and within-school streaming based on mathematics achievement, while controlling for social background (see Sect. 2.3). So we included students from two school contexts, here called *success context* and *at-risk context,* in our initial sample of the overarching intervention study (Prediger et al., 2022): First, a subsample **R'** of students at risk ($n = 323$) was selected among 1124 seventh-graders from 12 lower-tracked schools, and a subsample **S'** ($n = 266$) of successful students was selected among 279 sixth-graders from six higher-tracked schools (academic success operationalized by the higher track). In both subsamples, we selected students with fraction pretest scores below 15. While the subsample **R'** was selected based on weak achievement *after* completing the regular fraction teaching unit, the successful sixth-graders **S'** were selected before their systematic exposure to fractions. By varying age groups, we ensured the suitability of the intervention for both subsamples.

In the second sampling step, students in samples **R'** and **S'** were assigned to intervention and control groups in a cluster-randomized way. Of the intervention group with 394 students in 92 small groups with 3 to 6 students each, only 49 groups had the consent of all parents for video-recording the intervention sessions (20 groups of at-risk students and 29 groups of successful students). The students ($n = 210$) of these 49 groups form our video sample (Prediger et al., 2023).

The descriptive characteristics of the resulting video sample with its subsamples **R** (students at risk) and **S** (successful students) are listed in Table 1, which shows the video sample that was positively selected had slightly higher fraction scores and academic language proficiency.

The comparison of the subsamples **R** of students at risk and **S** of successful students in the video sample shows that as expected, the selected seventh graders at risk had higher fraction pretest scores than the successful sixth graders before their systematic encounter with fractions ($t(205) = -3.96$, $p < .001$). The samples do not differ in general cognitive ability ($t(204) = 1.72$, $p = .06$) or multilingual background ($\chi^2[1] = 0.23$, $p = .63$). However, sample **S** had a higher academic language proficiency ($t(205) = 3.77$, $p < .001$) and higher SES groups ($\chi^2[4] = 27.72$, $p < .001$).

3.5 Quality subdomains and quality features with different operationalizations

To analyze the interactional quality of the video data corpus of 49 teacher-led small groups taught with equal tasks, we conceptualized eight subdomains of interaction quality (Quabeck et al., 2023): talk-related activation (*TA*), talk-related participation (*TP*), conceptual activation (*CA*), conceptual participation (*CP*), discursive activation (*DA*), discursive participation (*DP*), lexical activation (*LA*), and lexical participation (*LP*). Following the variations in existing coding protocols (see Sect. 2.2), we distinguished (a) discursive richness, (b) conceptual richness, and (c) lexical richness.

Figure 2 depicts how eight conceptualized subdomains of enacted activation and individual students' participation were operationalized into 14 quality features with task-based, move-based, and/or practice-based operationalizations to follow the call for more transparency of operationalizations (Praetorius & Charalambous, 2018).

Talk-related subdomains serve as baseline (Sedova et al., 2019): *Talk-related participation* (*TP*) of individual students is measured by their relative individual talk time

as percentage of *time on task* (time for learning excluding social or classroom management interruptions, etc.). *Talk-related activation* (*TA*) is measured by class engagement, in other words, the sum of all students' relative talk times as percentage of time on task. For all other subdomains, the richness of the talk must be considered, and is operationalized in several ways, by the richness of tasks, teacher moves, or students' and teachers' co-constructed practices. Therefore, basic ratings were developed for conceptual, discursive, and lexical richness:

- The *tasks* were rated regarding their *conceptual richness* (following Kunter et al., 2013), *discursive richness* (discursively rich tasks demand students to explain meanings or report procedures), and *lexical richness* (lexically rich tasks explicitly promote lexical learning, e.g., by asking students to reflect on, collect, or use key phrases). The design of the intervention already included decisions about the tasks' intended conceptual, discursive, and lexical richness, thus the rating was low inferent. Based on this basic rating, we derived *task-based* operationalizations of quality features for interaction by capturing the length of interaction time spent on tasks of a particular degree of richness, for example, the task-based operationalization of quality feature *CA-t* in Fig. 4 is the relative length of the group's time spent on conceptually rich tasks (instead of procedural tasks).
- Teachers' *moves* were rated regarding their conceptual richness and lexical richness. *Conceptually rich moves* were identified as those that asked for, supported, or strengthened aspects of conceptual understanding, for example, when a teacher elicits ideas about equivalent fractions applied to equal soccer goal shooting rates (cf. Fig. 3). The time of the discussion elicited by this conceptually rich move was rated as move-based conceptually rich even if it took some turns before students answered the question. As discursively rich moves often do not elicit a rich discourse practice, no move-based operationalization of discursive richness was included. *Lexically rich moves* are those that explicitly promote lexical learning. For instance, in the example above, students might argue that "4 goals out of 5 attempts is as good as 8 out of 10." When a teacher takes this opportunity to introduce the phrase "equal share" and asks for students' re-formulations (e.g., "Yeah, we call it equal share, 4 out of 5 and 8 out of 10 are equally good. How can we rephrase what equal share means?"), the discussion following this move was considered lexically rich in the move-based operationalization. It can be very short if students do not elaborate on the move. Based on these basic ratings, *move-based quality features for the interaction* were derived by capturing the length of interaction time spent on moves of a particular degree of richness.

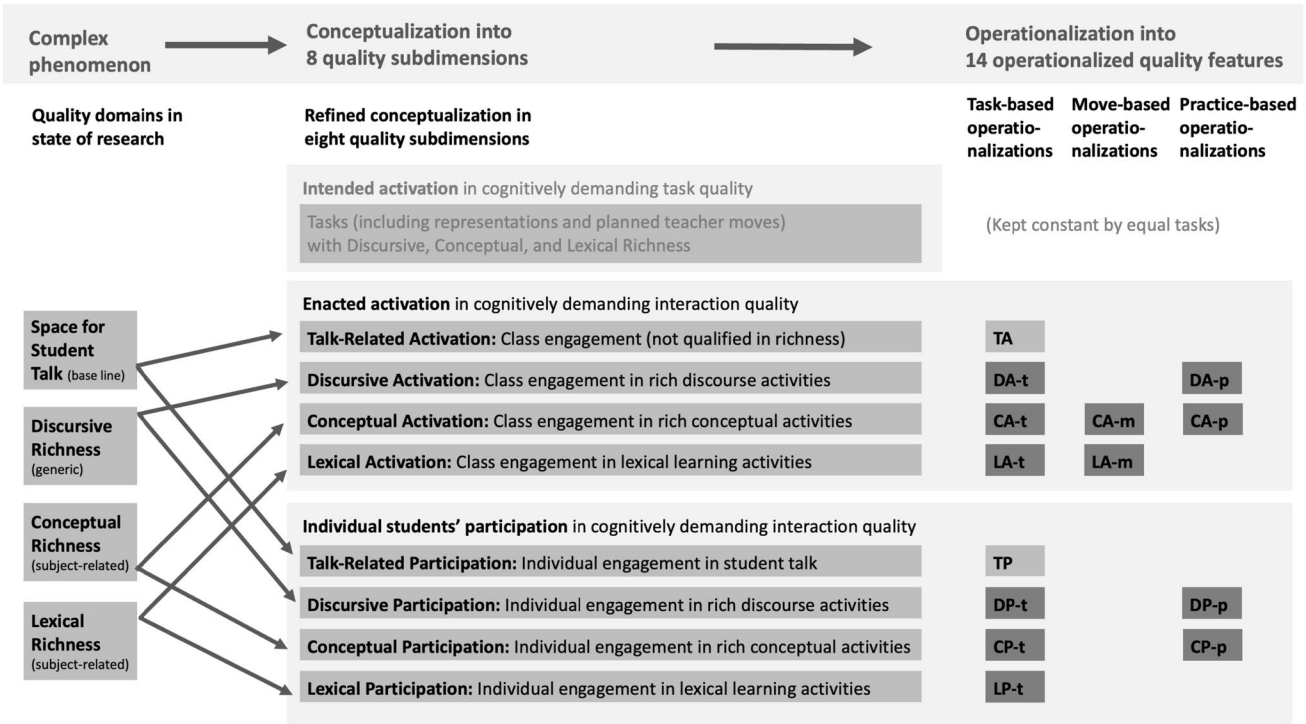

Fig. 2 Conceptualizing and operationalizing interaction quality in the enacted activation and individual students' participation (Quabeck et al., 2023)

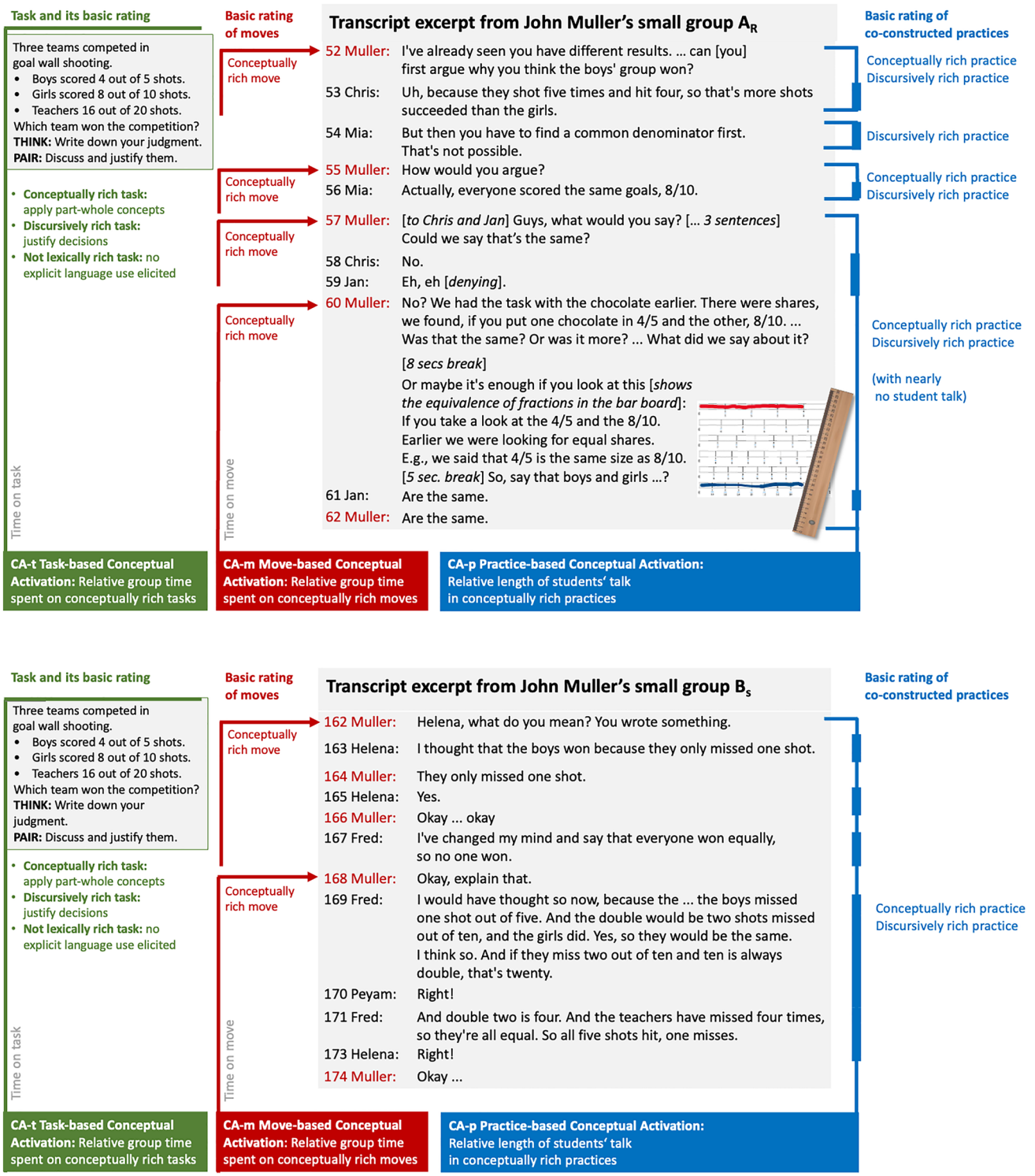

Fig. 3 Excerpts from two small groups and the basic rating of tasks, moves, and practices

- Finally, the practice-based operationalization required the most time-consuming basic rating, not only for the tasks and teacher moves, but for students' complete utterances. The utterances were rated with respect to the richness of the collectively established discourse *practices* they contributed to. A series of utterances was rated discursively rich when a discursively rich oral discourse practice such as elaborating an idea or reporting a pro-

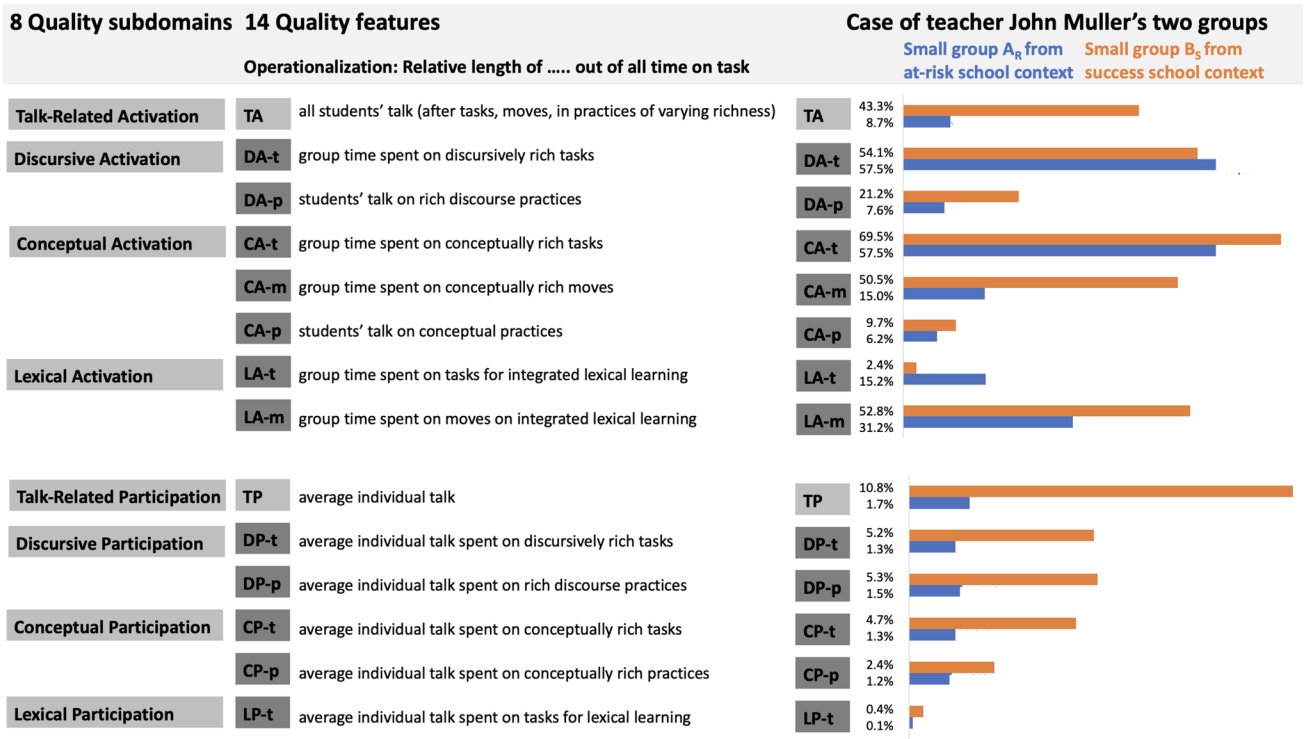

Fig. 4 Comparing 14 quality features of two small groups from different school contexts led by the same teacher

cedure was collectively established. For the example on goal shooting rates, students collectively explained how they approached the comparison of the shares: "First I draw a graphic representation of the shares and then I can compare them visibly." "Yes, then we see which share is bigger." This segment was not conceptually rich because students did not explain why the shares were equal. We rated those discourse *practices as conceptually rich* in which students explained meanings or described mathematical structures, for instance, "The shares are equal. Because when you have double attempts, you need double hits". This utterance explains the mathematical structure underlying the expansion by a factor of 2, so it was rated conceptually rich. Based on this basic rating, *practice-based quality features for the interaction* were derived by capturing the length of interaction time spent on particular discourse practices.

The highly inferent basic ratings (of conceptual and lexical moves and of practices) were conducted for a well-defined set of tasks (about order and equivalence of fractions, lasting 25 to 50 min). In total, 30 hours of video data were coded independently by two raters, with very good interrater reliabilities of Cohen's κ between 0.80 and 0.91.

With these ratings of conceptual, discursive, and lexical richness, every utterance of teachers and students was coded according to its richness with respect to the tasks currently discussed, the move it followed, or the established discourse practice.

To provide a unified measurement, all quality features were measured by time-related relative frequencies, which means by the percentage of talk time spent for a certain degree of richness (e.g., conceptual moves) in relation to total time on task (including times of silence).

In that way, the eight conceptualized subdomains of teachers' cognitively demanding and supportive enacted activation and individual students' participation were operationalized into 14 quality features with task-based, move-based, and/or practice-based operationalizations (Quabeck et al., 2023), as listed in Fig. 4.

3.6 Methods for the data analysis

To analyze how the quality features differ between school contexts (RQ1), we used Welch *t*-tests to compare the subsamples **R** and **S**, as it was more stable against variance heterogeneity (Delacre et al., 2017), and we determined Cohen's *d* for the effect size.

To analyze how the quality features are predicted by school context and students' individual prerequisites when all teachers share the same tasks (RQ2), we determined 14 hierarchical multiple regression models, one for each of the (highly correlating) quality features.

In each of the six regression models for individual participation ($n = 210$ students), we included one quality feature as the dependent variable and calculated to what extent they were predicted by school context, multilingual background, SES, general cognitive ability, and academic language proficiency (the last three as standardized metric variables, the first two by binary dummy variables). For the eight quality features of activation ($n = 49$ small groups), we included the average of these individual prerequisites as class composition variables.

Assumptions for conducting regression analysis were checked for each model (linearity, independent errors, normally distributed errors, homoscedasticity, and multicollinearity; Field, 2013). For each model, we tested if significant variance is explained, in total and in the quality features, for comparing effects across variables and models, we report the regression coefficient *b*.

4 Results

4.1 Case of John Muller in both school contexts

To increase the accessibility of later statistical results on differences between small groups (RQ1), we start by illustrating the meaning of the basic ratings and the quality features for the case of one teacher, pseudonymized John Muller, who taught one small group in success context (B_S) and one in at-risk context (A_R). Figure 3 shows two excerpts of his teacher-led group discussions with the basic rating of tasks, moves, and practices. From these ratings, time measurements are used to derive quality features, exemplified for CA-t, CA-m, CA-p.

John Muller worked with the same conceptually and discursively rich task and had prepared the same conceptually, discursively, and lexically rich teacher moves (exemplified in Fig. 3 for conceptually rich moves). However, the transcripts illustrate his stronger challenges to elicit substantial speech from the students in the at-risk context (i.e., more than two-word answers, as in Turn 58/59). So, he changed his follow-up moves to work with the resulting discourse practices, which varied in quality and in student contributions. In the success context, a single move elicits much richer practices, in the at-risk context, he finally gives the explanation himself.

In Fig. 4, we show the length comparisons of all tasks videorecorded in Muller's groups. Across all analyzed tasks, he established an interaction in which students talked 8.7% of the total time in small-group A_R from the at-risk context and 43.3% of the total time in small-group B_S from the success context (*talk-related activation*; *TA*). As a consequence, the *talk-related individual participation* also varied by a factor of 5. Beyond this surface structure (talk time notwithstanding its richness), the differences between John Muller's small groups varied for the operationalizations of richness.

Discursive activation can be operationalized by qualifying the relative length of group time spent on discursive tasks (*DA-t*). In *DA-t*, both small groups were very comparable (with 57.5% of the time in small-group A_R and 54.1% in small-group B_S, including John's talk time). However, much less time was devoted by the groups in engaging in rich discourse practices (such as arguing, explaining, or justifying decisions; *DA-p*). The difference between 7.6% in A_R and 21.2% in B_S in *DA-p* indicated that even when the tasks had discursively rich operators, students varied strongly in the proportion of the time that they really engaged in these requested rich discourse practices (instead of giving one-word answers).

The differences were even stronger for the individual *discursive participation*: The average relative length of time an individual student in small-group A_R tried to talk about discursively rich tasks was 1.3% in *DP-t*, while it was 5.2% in B_s. Similar differences occurred for the participation in rich discourse practices (*DP-p*).

For *conceptual activation*, small-group A_R invested less time in conceptually rich tasks than small-group B_S (57.5% vs. 69.7% for *CA-t*) did, yet the differences were much larger when operationalizing conceptual activation by the time spent after conceptually rich moves (15.0% vs. 50.5% for *CA-m*), whereas the groups' activation in conceptual practices (*CA-p*) revealed no relevant differences. In contrast, the differences in individual conceptual participation became significant (*CP-p*).

In these two groups, the teacher steered *lexical activation* less by the tasks (15.2% vs. 2.4% for *LA-t*) than by the moves (31.2% vs. 52.8% for *LA-m*), with a remarkably higher prevalence in the small-group B_S, so individual participation was also higher by a factor of 4.

This comparison of two cases of small groups raises the question of whether this is typical for the school contexts (analyzed in Sub Sect. 4.2) and whether this relates more to students' individual prerequisites or the school contexts (analyzed in Sub Sect. 4.3).

4.2 Prevalence of quality features in different school contexts

Research question RQ1 asks for differences of quality features in different school contexts. The descriptive results for the 14 quality features in 49 groups are documented in Table 2, for the whole sample and both subsamples from different school contexts.

Within the same subdomain, different task-based, move-based, and practice-based operationalizations detected

substantially different relative lengths. For example, the relative group time spent on conceptual tasks (*CA-t*, 77.4% of the time) was more than twice the time spent on teachers' conceptual moves (*CA-m*, 35.3%) or in established conceptual practices (*CA-p*, 23.5%). Thus, even for tasks with conceptual focus, this focus was not always reflected in teachers' moves, and teachers and students interactively co-constructed conceptual practices only in a third of the time. In contrast, although the proportion of time spent on tasks with explicit rich discursive task demands (*DA-t*) was only 11%, students' talk was part of rich discourse practices for 16.8% of the time (*DA-p*). Similarly, teachers' moves strengthened lexical activation: Only 10.7% of the time was dedicated to tasks that explicitly provided lexical learning opportunities (*LA-t*), but during 42.4%, the groups were lexically activated by teachers' lexical-integrating moves (*LA-m*).

The last two columns compare the differential learning conditions in at-risk contexts **R** and successful contexts **S**. They reveal comparable conceptual and lexical activation when operationalized by time spent on rich tasks (*CA-t* and *LA-t*) or lexical moves (*LA-m*), but significantly lower qualities for group **R** in at-risk contexts with respect to all other quality features.

4.3 Predicting quality features by individual prerequisites and school context

The focus of research question RQ2 is on the extent to which the quality features can be predicted by the school context and students' prerequisites (as individual prerequisites for quality features of individual participation and as average prerequisites in class composition for quality features of activation).

In Table 3, eight separate regression models for eight quality features of activation are documented, showing that almost none of the class composition variables significantly predicted the quality features of activation.

Only the percentage of multilingual students negatively predicted talk-related activation *TA* ($b = -.068$ for the relative length of student talk) and positively predicted move-based lexical activation *LA-m* ($b = .086$ for relative length of group time spent on lexically rich tasks). This means that when a group had only multilingual students, it spent approximately 8.6% more time on discussing after moves that could enhance lexical learning than in purely monolingual groups. In contrast, the school context predicted the quality features of talk-related activation ($b = .157$ for *TA*), practice-based discursive activation ($b = .077$ for *DA-p*), and move-based conceptual activation ($b = .101$ for *CA-m*): After controlling for class composition, groups from the success context spent an estimated 15.7% more time in student

Table 2 Distribution of enacted quality features with mean (and SD) for relative lengths: For 210 students in 49 small groups and differences between at-risk R and successful school contexts S (from Quabeck, 2023, Tables 8.2 and 8.3)

Quality feature		M (SD) in whole video sample	M (SD) in R in at-risk school contexts	M (SD) in S in successful school contexts	Welch *t* test for differences		
					t	*p*	Effect size d
Talk-related activation	TA	33.4% (14%)	23.1% (12%)	40.2% (10%)	**$t(37,52) = -5.57$**	**< .001**	**1.66**
Discursive activation	DA-t	11.0% (8%)	9.2% (6%)	12.1% (8%)	$t(46,31) = -1.36$	.09	0.38
	DA-p	16.8% (7%)	12.4% (5%)	19.7% (7%)	**$t(46,83) = -4.52$**	**< .001**	**1.24**
Conceptual activation	CA-t	77.4% (11%)	76.2% (10%)	78.2% (11%)	$t(44,00) = -0.83$	.205	0.24
	CA-m	35.3% (13%)	29.8% (11%)	39.0% (13%)	**$t(44,19) = -2.74$**	**.004**	**0.78**
	CA-p	12.1% (7%)	9.4% (5%)	14.0% (7%)	**$t(46,88) = -2.93$**	**.003**	**0.79**
Lexical activation	LA-t	10.7% (9%)	9.7% (8%)	11.4% (9%)	$t(44,17) = -0.57$	.29	0.16
	LA-m	42.4% (13%)	38.8% (10%)	44.8% (15%)	$t(46,88) = -1.43$	.08	0.39
Talk-related participation	TP	7.7% (5%)	5.4% (5%)	9.2% (5%)	**$t(178,83) = -5.49$**	**< .001**	**0.77**
Discursive participation	DP-t	1.3% (1%)	0.90% (1%)	1.6% (2%)	**$t(194,52) = -4.90$**	**< .001**	**0.67**
	DP-p	3.9% (3%)	2.8% (3%)	4.5% (3%)	**$t(197,09) = -4.06$**	**< .001**	**0.55**
Conceptual participation	CP-t	5.5% (4%)	4.0% (4%)	6.4% (4%)	**$t(183,07) = -4.49$**	**< .001**	**0.63**
	CP-p	2.8% (3%)	2.2% (2%)	3.2% (3%)	**$t(199,13) = -2.92$**	**.002**	**0.4**
Lexical participation	LP-t	0.9% (1%)	0.5% (1%)	1.2% (1%)	**$t(207,48) = -4.97$**	**< .001**	**0.64**

Significant predictors in bold letters

Table 3 Predictors for quality features of enacted activation: Eight linear regression models with school context and class composition variables (average learning prerequisites; from Quabeck, 2023, Tables 8.10–8.15)

	TA: Talk-related activation		DA-t: Discursive activation task based		DA-p: Discursive activation practice based		CA-t: Conceptual activation task based		CA-m: Conceptual activation move based		CA-p: Conceptual activation practice based		LA-t: Lexical activation task based		LA-m: Lexical activation move based	
	b	*SE*	*b*	*SE*	*b*	*SE*	*b*	*SE*	*b*	*SE*	*b*	*SE*	*b*	*SE*	*b*	*SE*
Intercept	.318**	.090	.715***	.120	.095	.058	.747**	.093	.197	.101	.140*	.056	.111	.078	.381**	.109
Success instead of at-risk context	**.157*** **	**.044**	.089	.059	**.077****	**.028**	.054	.045	**.101***	**.049**	.051	.027	.001	.038	.046	.053
Average fraction prior knowledge	− .002	.005	− .008	.006	− .001	.003	− .002	.005	− .003	.005	− .002	.003	.001	.004	− .0055	.006
Average language proficiency	− .0004	.003	.004	.004	.001	.002	.005	.003	.003	.003	.002	.002	− .003	.003	− .002	.004
Average socio-economic status	.019	.031	− .079	.042	− .004	.020	− .059	.032	− .030	.035	− .010	.019	.035	.027	− .002	.038
Average general cognitive abilities	− .007	.008	− .009	.011	− .001	.005	− .004	.008	.004	.009	− .006	.005	− .001	.007	.009	.010
Percentage multilingual students	**− .068***	**.033**	− .036	.044	.009	.021	− .013	.034	.066	.037	− .021	.021	.01	.029	**.086***	**.040**
R^2 / corrected R^2	.479/.405		.162/.042		.294/.193		.103/-.025		.243/.135		.187/.071		.056/.079		.184/067	
F(6, 42)	6.440		0.142		2.914		0.802		2.247		1.610		.417		1.573	
p	**< .001*****		.26		**< .05***		.57		**.057°**		.17		.86		.18	

Significance indicated on four levels: $^{\circ}p < 0.1$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

talk than those from at-risk contexts, 7.7% more time in rich discourse practices, and 10.1% more in talk initiated and supported by conceptual moves. In contrast, none of the variables predicted the time spent by small groups on a discursively or conceptually task (*DA-t, CA-t,* and *LA-t* had no significant predictors). In Table 4, six regression models for individual participation are documented, revealing that task-based conceptual participation ($b = -.013$ for *CP-t*) and practice-based conceptual participation ($b = -.012$ for *CP-p*) were predicted by students' multilingual background; and task-based lexical participation *LP-t* by language proficiency and socioeconomic status, albeit all with small regression coefficients. After controlling for individual prerequisites, the success school context was still significantly predictive for every quality feature, with small *b* (ranging from 0.007 for *LP-t* to 0.035 for *TP*).

5 Discussion

5.1 Embedding the findings into the state of research

The call for disentangling dimensions of instructional quality (Bostic et al., 2021; Mu et al., 2022) and for transparently accounting for operationalizations (Ing & Webb, 2012) is most important for subdomains of cognitive demand, the dimension treated in generic or subject-related ways (Praetorius & Charalambous, 2018; Brunner, 2018). In our project, we contribute to the disentanglement agenda with a particular focus on interaction quality while the task quality is systematically held constant (Quabeck et al., 2023; Prediger et al., 2023).

The analysis of *differential* interaction qualities is particularly interesting as most studies that *quantitatively* capture differential qualities have mainly referred to institutional effects through differential ambitions of curricula (operationalized e.g., by prioritized learning goals and quality of tasks; Oaks et al., 1992), whereas for interaction, mainly *qualitative* case studies have revealed large differences in learning opportunities shaped by low expectations in at-risk contexts (DIME, 2007). The few quantitative studies referring to differential class composition or school context effects in instructional quality have revealed incoherent findings with respect to differences in cognitive demands, depending on whether they were operationalized through task quality (Baumert et al., 2010) or observation of both, tasks and interaction (Fauth et al., 2021; Blömeke et al., 2022).

In our study, we could replicate existing qualitative observations of differential interaction qualities (DIME, 2007) and disentangle them into 14 quantified quality features suggested as operationalizations for subdomains of cognitive demand. The summary of comparisons and regression models in Figure 5 shows that *task-based operationalizations of activation* (*DA-t, CA-t,* and *LA-t*) were not significantly different; in other words, small groups from at-risk contexts or success contexts spent a similar

Table 4 Predictors for quality features of individual participation: Six linear regression models with school context and individual learning prerequisites (from Quabeck, 2023, Tables 8.4–8.9)

	TP: Talk-related participation		DP-t: Discursive participation task based		DP-p: Discursive participation practice based		CP-t: Conceptual participation task based		CP-p: Conceptual participation practice based		LP-t: Lexical participation task based	
	b	SE	*b*	SE	*b*	SE	*b*	SE	*b*	SE	*b*	SE
Intercept	.071**	.026	**.074****	**.019**	.015	.017	.051*	.021	.025	.014	.002	.006
Success context instead of at-risk context	**.035****	**.011**	**.024****	**.008**	**.017***	**.007**	**.027****	**.009**	**.013***	**.006**	**.007***	**.003**
Fraction prior knowledge	.0002	.002	− .002	.002	.0005	.002	.001	.002	.0001	.001	.001	.001
Language proficiency	− .0005	.001	.0001	.001	.0004	.001	.0001	.001	.0003	.0004	**− .001***	**.0002**
Socioeconomic status	.009	.007	− .002	.005	− .0005	.004	.003	.005	− .001	.004	**.004****	**.002**
General cognitive abilities	− .002	.002	− .002	.001	− .0003	.001	− .003	.001	− .001	.001	.001	.0004
Multilingual background	− .014	.007	**− .012***	**.005**	− .003	.005	**− .013***	**.006**	− .002	.004	.003	.002
R^2 / corrected R^2	.153/.127		.142/.116		.076/.049		.128/.101		.048/.0202		.157/.132	
F(6, 200)	6.004		5.527		2.760		4.873		1.699		6.203	
p	< .001***		< .001***		< .05*		0.001***		.12		.001**	

Significance indicated on four levels: *p < 0.05, *p < 0.01, ***p < 0.001

proportion of time on the given discursively, conceptually, or lexically rich tasks.

However, even in this laboratory situation with shared curriculum material and equal preparation of teachers in both contexts (in our laboratory context, which deviates from other studies), we see that the relative length of time that students talked in total (*TA*) after conceptual moves and in discursively and conceptually rich practices differed significantly (Table 2).

One might assume that these differences can be predicted by class composition or students' individual prerequisites. However, a main finding of this study is that after controlling for class composition, significant differences remained for *TA, DA-p*, and *CA-m* for generic and subject-related subdomains (Table 3). In small groups with more multilinguals, students took significantly less space to talk collectively (*TA*) and individually on discursively and conceptually rich tasks (*DP-t, CP-t*), but spent more time after moves initiating integrated lexical learning (*LA-m*).

Classroom culture also constrains students' individual opportunities for *participation* (all participation features were significantly predicted by the school context: *TP, DP-t, DP-p, Cp-t, CP-p,* and *LP-t*), even if co-constructively established. A higher chance of higher interaction quality is thus essentially related to *classroom cultural effects,* as substantiated here.

This means that learning opportunities through high interaction quality in mathematics classrooms can result from belonging to a learning milieu and possibly only additionally from individual and family learning preconditions. Indeed, the pretest and posttest assessment revealed that even after controlling for individual prerequisites and prior knowledge, adhering to the at-risk contexts predicted significantly lower learning gains (Prediger et al., 2022). With the current study, this predictive power can also be explained by differential interaction qualities.

In total, these findings call for focusing deep structures of interaction (not surface structures of talk time). Differential interaction qualities (in conceptual, discursive, and lexical richness) could be identified through a research design that deliberately excluded simple institutional effects (of curriculum, task, and teacher preparation) and controlled for prerequisite effects. In this way, we extend the list of documented (class composition, prerequisite, and institutional) effects to include *classroom cultural effects*, beyond teachers' often documented low expectations for at-risk students (DIME, 2007): While Cai et al. (2020) called for asking "How does teaching contribute to creating and realizing learning opportunities?" (p. 19), we provide quantitative evidence that learning opportunities in classrooms are not only led by *teaching*, but are interactively established with what students bring into the discussion. These findings resonate with general case study findings showing that interaction is always co-constructed by teachers *and* students (Howe & Abedin, 2013; Walshaw & Anthony, 2008), but from now on, we should take into account that beyond students' individual capabilities and teachers' (low expectation-shaped) choices of tasks and moves, the co-constructively established classroom cultures can be heavily influenced by school contexts, even in our laboratory when the curriculum and teachers are the same (Fig. 4).

This finding emphasizes the need for further discussion about early between-school tracking, particularly in the context of equity and reproduction of educational inequality.

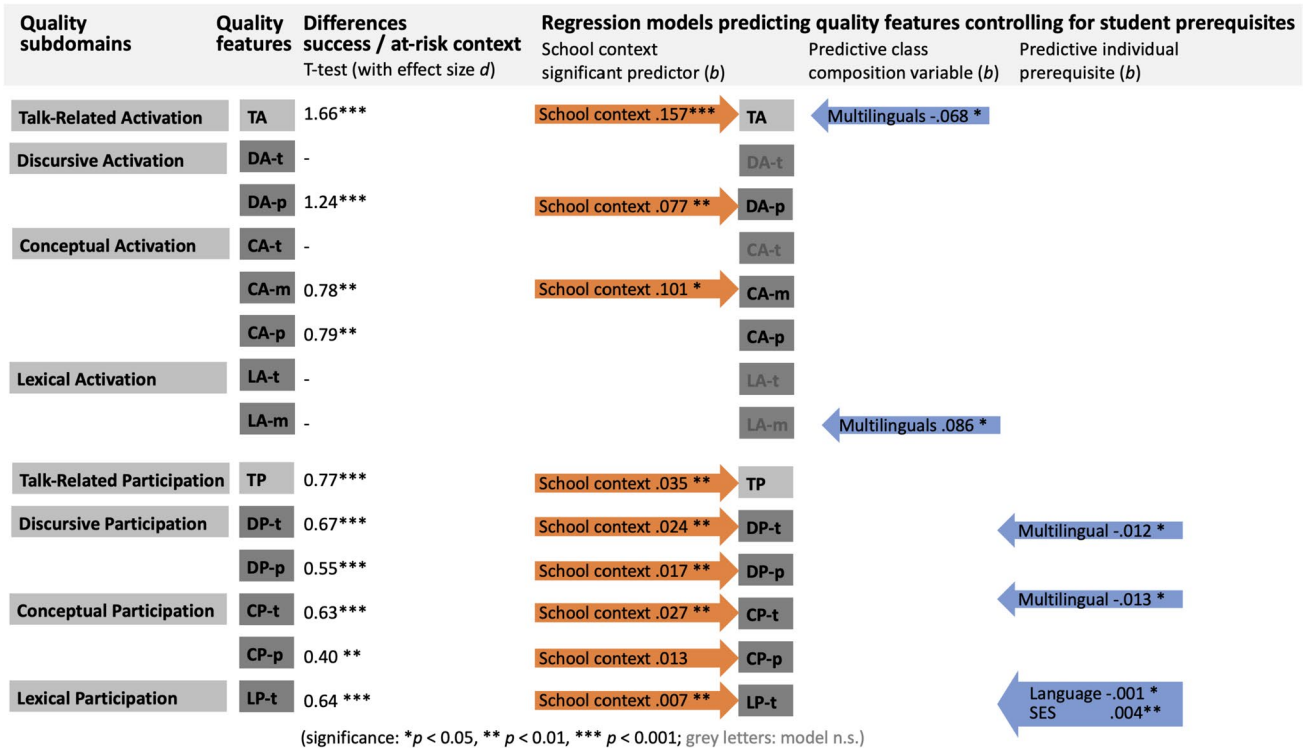

Fig. 5 Summary of differential findings: Influence of school contexts, class composition, and individual prerequisites on interaction qualities

Because even when the best teachers and curriculum material are available in at-risk contexts, the co-constructed nature of interaction quality might limit students' learning opportunities.

Finally, we repeat that the differences became observable through move-based and practice-based operationalizations of activation. Thus, our study also highlights the importance of operationalizing interaction quality beyond the richness of tasks.

5.2 Methodological limitations and methodological learnings

Of course, the study must be interpreted against the background of its methodological limitations, starting with operationalizations of "at-risk context," which is tied to the particular German context (with early between- and in-school tracking) and of socio-economic status (by the book-at-home index, which will become invalid soon, due to digitalization), that will require other operationalizations when similar investigations are transferred to other contexts.

We do not claim to have fully covered all existing generic and subject-related conceptualizations of cognitive demands (Mu et al., 2022), but have restricted our analysis to a particular selection of generic and subject-related conceptualizations in order to be able to systematically vary the operationalization. The measurement in time-related relative frequencies allowed us to design all operationalizations in comparable ways, but future studies should analyze whether the results might be different with other operationalizations and measurements.

For the moment, the findings are tied to the data gathering context of 49 small groups, but in the future, whole-class studies should also be conducted, even if the subtle classroom cultural effects of teachers' facilitation might be more difficult to grasp. To quantitatively capture individual students' participation and compare these values between school contexts, the small groups have already provided somewhat fragile data with relative length of individual participation ranging from 0.5% to 9.2% of the time. We must therefore suspect that these quality features have limited stability across sessions, which should be analyzed in the future.

In addition, future research should overcome this study's limitations of measuring participation only through active verbal contributions, as other studies have deconstructed the simple connection between active participation in classroom talk and mathematics achievement, and pointed to learning through silent participation (O'Connor et al., 2017). As it may be difficult to adequately capture silent participation through video coding, future research should find modern technologies such as eye movement to capture participation in broader terms.

In its current state, though, this study already substantially contributes to the methodological discourse in research on instructional quality, showing the high relevance of exactly accounting for the theoretical and methodological decisions

on conceptualizations and operationalizations (Ing & Webb, 2012; Mu et al., 2022), which change their prevalence substantially with school context: when measuring times after particular tasks, we can substantially overestimate the intended activation and participation in underprivileged school contexts, as the time really spent in rich practices has much larger differences between the school contexts than the time spent in tasks. Thus, the effort in different conceptualizations and operationalizations allows bridging this quantitative study to insights from qualitative and even highly interpretative classroom studies (Walshaw & Anthony, 2008; DIME, 2007).

Acknowledgements The project MuM-MESUT (Developing Conceptual Understanding by Language Support) was funded by the German Research Foundation (DFG Grants No. PR 662/14-2 to S. Prediger and ER 880/3-3 to K. Erath).

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations

Conflict of interest No potential conflict of interest was reported by the authors.

References

Baumert, J., Kunter, M., Blum, W., Brunner, M., Voss, T., Jordan, A., & Tsai, Y.-M. (2010). Teachers' mathematical knowledge, cognitive activation in the classroom, and student progress. *American Educational Research Journal, 47*(1), 133–180. https://doi.org/10.3102/0002831209345157

Becker, M., Kocaj, A., Jansen, M., Dumont, H., & Lüdtke, O. (2022). Class-average achievement and individual achievement development. *Journal of Educational Psychology, 114*(1), 177–197. https://doi.org/10.1037/edu0000519

Blömeke, S., Jentsch, A., Ross, N., Kaiser, G., & König, J. (2022). Opening up the black box: Teacher competence, instructional quality, and students' learning progress. *Learning and Instruction, 79*(101600), 1–11. https://doi.org/10.1016/j.learninstruc.2022.101600

Bostic, J., Lesseig, K., Sherman, M., & Boston, M. (2021). Classroom observation and mathematics education research. *Journal of Mathematics Teacher Education, 24*(1), 5–31.

Boston, M. (2012). Assessing instructional quality in mathematics. *The Elementary School Journal, 113*(1), 76–104. https://doi.org/10.1086/666387

Brophy, J. (2000). *Teaching* (Educational Practices Series Vol 1). Int Academy of Education.

Brühwiler, C., & Blatchford, P. (2011). Effects of class size and adaptive teaching competency on classroom processes and academic outcome. *Learning and Instruction, 21*(1), 95–108. https://doi.org/10.1016/j.learninstruc.2009.11.004

Brunner, E. (2018). Qualität von Mathematikunterricht: Eine Frage der Perspektive [Quality of mathematics instruction: A question of perspective]. *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

Delacre, M., Lakens, D., & Leys, C. (2017). Why psychologists should by default use Welch's t-test instead of student's t-test. *International Review of Social Psychology, 30*(1), 92–101. https://doi.org/10.5334/irsp.82

DIME–Diversity in Mathematics Education Center for Learning and Teaching (2007). Culture, race, power and mathematics education. In: F. Lester (Ed.) *Second handbook of research on mathematics teaching and learning*. Information Age. pp. 405–433

Fauth, B., Atlay, C., Dumont, H., & Decristan, J. (2021). Does what you get depend on who you are with? Effects of student composition on teaching quality. *Learning and Instruction, 71*(101355), 1–9. https://doi.org/10.1016/j.learninstruc.2020.101355

Field, A. (2013). *Discovering statistics using IBM SPSS statistics*. Sage.

Flanders, N. A. (1970). *Analyzing teaching behavior*. Addison-Wesley.

Gibbons, P. (2002). *Scaffolding language, scaffolding learning*. Heinemann.

Helmke, A. (2009). *Unterrichtsqualität und Lehrerprofessionalität [Instructional quality and teacher professionalism]*. Kallmeyer.

Hill, H. C., Blunk, M. L., Charalambous, C. Y., Lewis, J. M., Phelps, G. C., Sleep, L., & Ball, D. L. (2008). Mathematical knowledge for teaching and the mathematical quality of instruction. *Cognition and Instruction, 26*(4), 430–511. https://doi.org/10.1080/07370000802177235

Howe, C., & Abedin, M. (2013). Classroom dialogue: A systematic review across four decades of research. *Cambridge Journal of Education, 43*(3), 325–356. https://doi.org/10.1080/0305764X.2013.786024

Ing, M., & Webb, N. M. (2012). Characterizing mathematics classroom practice: Impact of observation and coding choices. *Educational Measurement: Issues and Practice, 31*(1), 14–26. https://doi.org/10.1111/j.1745-3992.2011.00224.x

Kunter, M., Baumert, J., Blum, W., Klusmann, U., Krauss, S., & Neubrand, M. (Eds.). (2013). *Cognitive activation in the mathematics classroom and professional competence of teachers*. Springer.

Lampert, M., & Cobb, P. (2003). Communication and language. In J. Kilpatrick & D. Schifter (Eds.), *A research companion to principles and standards for school mathematics* (pp. 237–249). NCTM.

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

Lipowsky, F., Rakoczy, K., Pauli, C., Reusser, K., & Klieme, E. (2007). Gleicher Unterricht–gleiche Chancen für alle? [Equal instruction – equal opportunities for all?]. *Unterrichtswissenschaft, 35*(2), 125–147.

Maaz, K., Trautwein, U., Lüdtke, O., & Baumert, J. (2008). Educational transitions and differential learning environments. *Child*

Development Perspectives, 2(2), 99–106. https://doi.org/10.1111/j.1750-8606.2008.00048.x

Moschkovich, J. (2015). Academic literacy in mathematics for English learners. *The Journal of Mathematical Behavior, 40*, 43–62. https://doi.org/10.1016/j.jmathb.2015.01.005

Mu, J., Bayrak, A., & Ufer, S. (2022). Conceptualizing and measuring instructional quality in mathematics education: A systematic literature review. *Frontiers in Education, 7*(994739), 1–30. https://doi.org/10.3389/feduc.2022.994739

Neubrand, M., Jordan, A., Krauss, S., Blum, W., & Löwen, K. (2013). Task analysis in COACTIV. In M. Kunter, J. Baumert, W. Blum, U. Klusmann, S. Krauss, & M. Neubrand (Eds.), *Cognitive activation in the mathematics classroom and professional competence of teachers.* Springer.

Ni, Y., Zhou, D.-H.R., Cai, J., Li, X., Li, Q., & Sun, I. X. (2018). Improving cognitive and affective learning outcomes of students through mathematics instructional tasks of high cognitive demand. *Journal of Educational Research, 111*(6), 704–719. https://doi.org/10.1080/00220671.2017.1402748

O'Connor, C., Michaels, S., Chapin, S., & Harbaugh, A. G. (2017). The silent and the vocal: Participation and learning in whole-class discussions. *Learning and Instruction, 48*, 5–13. https://doi.org/10.1016/j.learninstruc.2016.11.003

Oakes, J., Gamoran, A., & Page, R. (1992). Curriculum differentiation: Opportunities, outcomes, and meanings. In P. Jackson (Ed.), *Handbook of research on curriculum* (pp. 570–608). Macmillan.

OECD (2020) *Global Teaching InSights: A Video Study of Teaching*. OECD.

Pauli, C., & Reusser, K. (2015). Discursive cultures of learning in (everyday) mathematics teaching. In L. B. Resnick, C. S. C. Asterhan, & S. N. Clarke (Eds.), *Socializing intelligence through academic talk and dialogue.* AERA.

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

Praetorius, A.-K., & Charalambous, C. Y. (2018). Classroom observation frameworks for studying instructional quality. *ZDM—Mathematics Education, 50*(3), 533–553. https://doi.org/10.1007/s11858-018-0946-0

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

Prediger, S., Erath, K., Weinert, H., & Quabeck, K. (2022). Only for multilingual students at risk? Cluster-randomized trial on language-responsive instruction. *Journal for Research in Mathematics Education, 53*(4), 255–276. https://doi.org/10.5951/jresematheduc-2020-0193

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

Quabeck, K., Erath, K., & Prediger, S. (2023). Measuring interaction quality in mathematics instruction. *Journal of Mathematical Behavior, 70*(101054), 1–17. https://doi.org/10.1016/j.jmathb.2023.101054

Quabeck, K. (2023). *Interaktionsqualität im sprachbildenden Mathematikunterricht* [Interaction quality in language-responsive mathematics classrooms]. PhD thesis. TU Dortmund University (to be published by Springer in 2024).

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*, 491–506.

Sedova, K., Sedlacek, M., Svaricek, R., Majcik, M., Navratilova, J., Drexlerova, A., Kychler, J., & Salamounova, Z. (2019). Do those who talk more learn more? *Learning and Instruction, 63*(101217), 1–11. https://doi.org/10.1016/J.LEARNINSTRUC.2019.101217

Spreitzer, C., Hafner, S., Krainer, K., & Vohns, A. (2022). Effects of generic and subject-didactic teaching characteristics on student performance in mathematics in secondary school A scoping review. *European Journal of Educational Research, 11*(2), 711–737. https://doi.org/10.12973/eu-jer.11.2.711

Stigler, J. W., Gonzales, P., Kawanaka, T., Knoll, S., & Serrano, A. (1999). *The TIMSS Videotape Classroom Study*. National Center for Education Statistics.

Walshaw, M., & Anthony, G. (2008). The teacher's role in classroom discourse: A review of recent research into mathematics classrooms. *Review of Educational Research, 78*(3), 516–551. https://doi.org/10.3102/0034654308320292