International Journal of Science and Mathematics Education (2024) 22:885–909
https://doi.org/10.1007/s10763-023-10389-4

Effects of Interaction Qualities Beyond Task Quality: Disentangling Instructional Support and Cognitive Demands

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

Received: 16 September 2022 / Accepted: 20 May 2023 / Published online: 1 August 2023

Abstract

Instructional quality dimensions of cognitive demands and instructional support have been shown to have an impact on students' learning gains. Existing operationalizations of these dimensions have mostly used comprehensive ratings that combine various subdimensions of task quality and interaction quality. The current study disentangles interaction quality in a video data corpus study (of 49 middle school classrooms sharing the same tasks) to identify those quality features that predict students' learning gains in conceptual understanding. The regression analysis reveals that quality features of students' individual engagement do not predict individual student learning, whereas teachers' support of learning content-relevant vocabulary predicts the small groups' learning. For at-risk students, the collective time spent on conceptual practices (i.e. explaining meanings of concepts) on students' learning is significantly predictive. The observation that different operationalizations (for similar aspects of interaction quality) lead to different impacts on the learning gains contributes to ongoing research efforts to refine and increase insight into aspects of interaction quality.

Keywords Cognitive demands · Conceptual richness · Interaction quality · Lexical support

✉ Susanne Prediger
prediger@math.tu-dortmund.de

[1] IPN Leibniz Institute for Science and Mathematics Education, Hausvogteiplatz 5-7, 10017 Berlin, Germany

[2] TU Dortmund University, Vogelpothsweg 87, 44221 Dortmund, Germany

[3] Martin Luther University Halle-Wittenberg, 06099 Halle (Saale), Germany

Introduction: Need to Disentangle the Quality of Cognitively Demanding and Supportive Interaction

International research on subject-matter teaching has repeatedly shown that students' learning gains are substantially influenced by the quality of instruction (Brophy, 2000; Cai et al., 2020; Hiebert & Grouws, 2007). Two recent research reviews on instructional quality frameworks revealed that in particular the quality dimensions of *cognitive demands* and *instructional support* had strong effects on students' learning gains, two dimensions with strong overlaps, and multiple conceptualizations and operationalizations (Praetorius & Charalambous, 2018; Spreitzer et al., 2022). The authors of the TALIS study therefore conceded that "we are only beginning to understand what makes a difference in terms of quality teaching" (Organization for Economic Cooperation and Development [OECD], 2020, p. 14) and called for further striving for depth in the ways instructional quality is measured and then related to students' learning gains.

Cognitive demands are defined as the extent to which teachers "create and maintain an environment of productive intellectual challenge that is conducive to students' mathematical development … [and support] students in grappling with … concepts" (Schoenfeld, 2014, pp. 407–408). Cognitive demands are often operationalized by the cognitive richness of the tasks that teachers launch (e.g. Kunter et al., 2013; Ni et al., 2018). However, Henningsen and Stein (1997) had already shown that even with high-quality tasks, cognitive demands can be maintained, increased, or reduced in the *interaction* between teachers and students. Similarly, the quality dimension *instructional support* (operationalized by Pianta & Hamre, 2009) is characterized by the suitability of tasks used to enhance students' concept development, but also by aspects of interaction, e.g. the microscaffolding that teachers provide (Allen et al., 2013).

We therefore focus on *classroom interaction*, defined as ways in which teachers and students communicate to interactively establish shared meanings and discuss mathematical ideas (Bauersfeld, 1988; Krummheuer, 2011). Qualitative studies have shown how strongly students' learning opportunities tend to be shaped by classroom interaction (Walshaw & Anthony, 2008). In *quantitative* instructional quality frameworks, task quality and interaction quality are usually considered together in comprehensive ratings, but have not yet been scrutinized into their separate contributions (Howe et al., 2019). In this paper, we aim at further disentangling the possible effects of interaction quality on students' learning gains. To separate the effects of interaction quality features from effects of task quality, we chose a video data corpus in which the task quality was kept constant as all observed teachers used the same tasks, so that we could zoom into the interaction quality and pursue the following research question:

How do different features of interaction quality impact students' learning gains?

In the theory section, we summarize the state of research on interaction quality on which our conceptual framework is based. The methodology section presents the data corpus and the methods for quantitative in-depth video coding and statistical analysis. The results section reveals the empirical findings, showing which quality features are predictive for middle school students' learning gains in the mathematical topic in view: conceptual understanding of fractions.

Conceptual Framework for Capturing Cognitively Demanding and Supportive Interaction Quality Beyond Task Quality

Whereas early quantitative coding protocols captured surface structures of instruction such as teachers' and students' talk time (Flanders, 1970) or activity structures (group work, seat work, and whole-class discussion; Stigler et al., 1999), later coding protocols successively captured deeper structures. Within the heterogeneous coding protocols for deeper structures (Praetorius & Charalambous, 2018; Spreitzer et al., 2022), three main dimensions have often reoccurred: *classroom management* (dealing with behavioral disruptions and optimizing time on task), *socioemotional support* (encouragement and social climate), and a third dimension concerning the instruction in a narrower sense, conceptualized broadly as *instruction* (OECD, 2020), as *cognitive demands* (Lipowsky et al., 2009; Neugebauer & Prediger, 2023; Ni et al., 2018; Praetorius et al., 2018; Schoenfeld, 2014) or (with a slightly different focus) as *instructional support* (Allen et al., 2013; Pianta & Hamre, 2009). It has mainly been this third dimension of *cognitive demands and instructional support* that has been disentangled in recent subject-specific coding protocols for instructional quality (surveyed by Bostic et al., 2021), with some subject-related aspects of socioemotional support such as dealing with errors (Spreitzer et al., 2022). With our study, we contribute to disentangling the third dimension of content-related instruction (cognitive demands and instructional support), with a particular focus on interaction quality, and we show which quality feature really predicts learning gains in our data.

Existing Research on Cognitive Demands

Cognitive demands (also called "cognitive activation"; Lipowsky et al., 2009; Ni et al., 2018; Praetorius et al., 2018) count as a key dimension of instructional quality. Although a container construct with many different conceptualizations and operationalizations (as problematized by Praetorius & Charalambous, 2018; Spreitzer et al., 2022), a shared theoretical idea taken from teaching–learning theories has developed that has emphasized the need for higher-order thinking (such as conceptual understanding or mathematical reasoning practices) and students' deep and targeted engagement in intense thinking processes (Hiebert & Grouws, 2007). Quantitative studies on instructional quality have revealed a measurable impact of this dimension of cognitive demands on students' mathematical learning gains (e.g. Kunter et al., 2013; Lipowsky et al., 2009; Ni et al., 2018; Praetorius et al., 2018). These findings have been robust for heterogeneous conceptualizations, but have seemed to vary between students at risk and successful students (Bostic et al., 2021; Cai et al., 2020).

To further unpack the heterogenous conceptualizations, we disentangle them into various components of instruction identified as shaping what students can learn:

> The emphasis teachers place on different *learning goals*, … the kinds of *tasks* they pose, the kinds of *questions they ask* and responses they accept, the nature of the *discussions* they lead—all are part of teaching and all influence the opportunities … to learn (Hiebert & Grouws, 2007, p. 379, *italics* added).

The conceptualizations of cognitive demands in quantitative coding protocols have tended to differ in the components that have been prioritized: teachers' focus on *learning goals* (higher-order thinking learning goals such as conceptual understanding or mathematical reasoning practices; e.g. Kunter et al., 2013; Ni et al., 2018); the cognitive level of *tasks* (Decristan et al., 2015; Kunter et al., 2013; Ni et al., 2018; Stein & Lane, 1996; Stigler et al., 1999); the kinds of teacher *moves*, in other words, the questions asked and responses accepted (Decristan et al., 2015; Hsu et al., 2023; Lipowsky et al., 2009; OECD, 2020; Schlesinger et al., 2018); and the nature of *discourses* supporting concept development and productive struggle (Howe et al., 2019; Lipowsky et al., 2009; Neugebauer & Prediger, 2023; OECD, 2020). Some studies have captured mainly the *task quality* (Kunter et al., 2013; Ni et al., 2018), although the relevance of teacher-student interaction to maintain, expand, or reduce the cognitive demands has often been documented (from Henningsen & Stein, 1997, to Zhou et al., 2023).

When capturing *interaction quality beyond the task quality* and teachers' moves, students' engagement needs to be considered. Early operationalizations focused simply on students' space to talk (Flanders, 1970), but talk time alone has been shown to not be quantitatively predictive for learning gains (Inagaki et al., 1998; Pauli & Lipowsky, 2007). Many case studies, in turn, have specified conditions of mathematical richness and discursive richness of the talk (Lampert & Cobb, 2003; Walshaw & Anthony, 2008), in other words, the conceptual depth of the initiated mental processes and the complexity of elicited and supported discourse practices such as explaining or arguing. Each of these aspects of richness has also seemed to depend on the school contexts, with higher-tracked schools or schools in privileged areas providing richer learning opportunities than lower-tracked schools (Bostic et al., 2021; Cai et al., 2020; Pauli & Reusser, 2015), so school contexts seem to matter.

In quantitative coding protocols for instructional quality, mathematical richness and discursive richness have been measured with heterogeneous conceptualizations and operationalizations, which has raised concerns about missing coherence and limited transparency (Bostic et al., 2021; Cai et al., 2020; Praetorius & Charalambous, 2018). Additionally, many coding protocols were criticized to capture mainly teachers' moves for activating students, while neglecting the *richness of students' real participation* beyond talk time (Howe et al., 2019; Pauli & Reusser, 2015). However, teachers' provided activation and individual students' participation are structurally different phenomena (Brühwiler & Blatchford, 2011; Helmke, 2009). For example, some coding protocols have measured the mathematical richness almost exclusively by the demands of teacher moves (TEDS, Schlesinger et al., 2018) without considering students' answers, whereas others have captured mathematical richness in the interplay of teachers' prompts and students' contributions or the length of student contributions (e.g. in MQI, Hill et al., 2008; TALIS, OECD, 2020; or PYTHAGORAS, Pauli & Reusser, 2015). In most rating protocols, teachers' provided activation and individual students' participation have been mixed (e.g. MQI, Hill et al., 2008) and not even clearly distinguished in the papers' theoretical parts (as criticized by Praetorius & Charalambous, 2018), although these components were distinct in the supply-use model for classroom research (Brühwiler & Blatchford, 2011; Helmke, 2009).

Existing Research on Instructional Support

The quality dimension of *instructional support* was promoted by Pianta and Hamre (2009) in their CLASS protocol. Instructional support has substantial overlaps with cognitive demands in other coding protocols in that it covers the subdimensions of content understanding (e.g. "teacher presentation of content within a broader intellectual framework"), analysis and problem solving (e.g. "emphasis upon engaging students in highe-order thinking skills"), and quality of feedback (e.g. "provision of contingent feedback designed to challenge students and expand their understanding of a concept"; Allen et al., 2013). The focus has not been only on what to request from students but also on how to support students to meet these demands:

> Instructional supports do not focus solely on the content of curriculum or learning activities, but rather on the ways in which teachers implement these to effectively support cognitive and academic development. Teachers who … give consistent, timely, and process-oriented feedback; and work to extend students' language skills tend to have students who make greater achievement gains. (Pianta & Hamre, 2009, p. 113).

One often addressed aspect of in-the-moment support in mathematics classrooms has been the use of multiple representations (Hill et al., 2008; Schlesinger et al., 2018; Schoenfeld, 2014). In contrast, language support (a relevant subdimension of instructional support by Pianta & Hamre, 2009), in particular lexical support, is a relevant additional aspect that is not sufficiently covered by other subdimensions of cognitive demands. Teachers' rich lexical support for students' vocabulary acquisition has proven effective for literacy acquisition (Carlisle et al., 2013; Kohlmeier, 2018), and has also been shown to be productive in mathematics classrooms when embedded in rich discourse practices such as explaining meanings or arguing (Gibbons, 2002; Moschkovich, 2015; Smit et al., 2013).

As with cognitive demands, the existing coding protocols for instructional support holistically combine teachers' actions with the implemented student participation, so this requires further disentanglement in our conceptual framework.

Conceptual Framework Disentangling Activation and Participation in Subdimensions of Cognitively Demanding and Supportive Interaction Quality

The documented general need for further disentangling subdimensions of instructional quality (OECD, 2020; Praetorius & Charalambous, 2018) particularly applies to the overlapping dimensions being examined in this paper: cognitive demand and instructional support. Whereas cognitively demanding *task quality* has been captured in depth (e.g. Kunter et al., 2013), *interaction quality* needs to be further disentangled, in particular for classrooms in which high task quality is achieved (Bostic et al., 2021; OECD, 2020; Pauli & Reusser, 2015).

Our disentanglement of cognitively demanding and supportive interaction quality (Quabeck et al., 2023) draws upon the theoretical and methodological framework of the *supply-use model* (Brühwiler & Blatchford, 2011; Helmke, 2009; Vieluf et al., 2020), in which instruction is investigated with respect to *teachers' supply* and *students' individual use*. Given that interaction is co-constructed by teachers and students (Bauersfeld, 1988; Vieluf et al., 2020; Walshaw & Anthony, 2008), we adapted the supply-use model by splitting teachers' supply into subdimensions that capture (1) teachers' intended activation (task quality and teachers' planned moves, without their uptake in the co-constructed interaction) and (2) teachers' activation of the whole class as enacted in the interaction. Students' use is conceptualized as (3) the participation of individual students in this interaction.

Furthermore, we distinguish four quality domains that have been identified as distinct and relevant in qualitative case studies on interaction (Lampert & Cobb, 2003; Walshaw & Anthony, 2008): (a) space for student talk, (b) conceptual richness, (c) discursive richness, and (d) lexical richness. By these three different versions of richness, we aim to unpack the container construct richness for which the literature review reveals various distinctions (see below).

Table 1 articulates the conceptualization for each of the three subdimensions from the adapted supply-use model. As our focus is on *interaction*, we will continue later with the second and third column. We capture space for student talk as a baseline for more qualified dimensions of richness and narrow mathematical richness down to conceptual richness.

Recent research reviews on instructional quality frameworks (Praetorius & Charalambous, 2018; Spreitzer et al., 2022) and more refined coding protocols for interaction quality (Quabeck et al., 2023) have revealed that for each of these nine conceptualizations from Table 1, diverse operationalizations can be found. Besides the still prevailing questionnaires for teachers and students (Spreitzer et al., 2022, reported them in nearly half of the reviewed studies), existing observational coding protocols work with at least three operationalizations of interaction quality:

- *task-based* (rating the richness of demands and supports through the tasks),
- *move-based* (rating the quality of demands and supports posed by teachers' moves), and
- *practice-based operationalizations* (rating the quality of collectively or individually enacted practices emerging in the interaction).

Table 1 Different conceptualizations of cognitively demanding and supportive interaction quality in teachers' intended activation (*later held constant*), enacted activation, and individual students' participation

Quality domains	*Teachers' intended activation*	Enacted activation in the interaction	Individual students' participation
Space for student talk	*Offered space for student talk (e.g. by teachers' questions)*	Class engagement (space for all student talk)	Individual participation in student talk
Conceptual richness	*Conceptual and other high cognitive demands and supports (e.g. by tasks, representations, and teachers' moves)*	Class engagement in rich conceptual activities (e.g. in a conceptual task, after a conceptual move, in a conceptual practice)	Individual participation in rich conceptual activities
Discursive richness	*High discursive demands and supports (e.g. by tasks, representations, and teachers' moves)*	Class engagement in rich discourse activities (e.g. explaining) or referencing to each other	Individual participation in rich discourse activities (e.g. explaining)
Lexical richness	*Lexical support for students' lexical learning focus*	Class engagement in lexical learning activities	Individual engagement in lexical learning activities

For example, some operationalizations of discursive richness have been criticized as too simplifying (Pauli & Reusser, 2015), when teachers' intended activation is captured by move-based measures (e.g. TIMSS, Hsu et al., 2023; Stigler et al., 1999) instead of enacted classroom practices. Other studies have utilized a combination of moves and practices (e.g. MQI or IQA, as cited in Bostic et al., 2021), or have mainly focused on practices (e.g. students' and teachers' explanations in TALIS, OECD, 2020) for capturing teachers' enacted discursive richness in the interaction. Students' 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. PYTHAGORAS, Lipowsky et al., 2009). These task-based, move-based, or practice-based operationalizations still vary with respect to the rating or coding, they are either rated roughly in time segments of different sizes (Praetorius & Charalambous, 2018), or measured in turn-related (e.g. classroom discourse in PYTHAGORAS, Lipowsky et al., 2009), sentence-related (e.g. number of words spoken in TIMSS video study, Stigler et al., 1999), or time-related frequencies (e.g. Sedova et al., 2019).

In a preliminary study, we showed that when these different task-based, move-based, and practice-based operationalizations are systematically compared, they mostly have only weak correlations, so they seem to measure different phenomena (Quabeck et al., 2023). In this paper, we investigate these different conceptualizations and operationalization for the first time with respect to their predictive power for students' learning gains.

Methodological Framework for Scrutinizing Interaction Quality

Overall Research Design

After the theory section, we can refine the research question as follows:

How do the quality features of cognitively demanding and instructionally supportive interaction impact students' learning gains when all teachers share the same tasks and representations?

The research question is pursued within the framework of the supply-use model (Brühwiler & Blatchford, 2011; Helmke, 2009; Vieluf et al., 2020), which we adapted in Fig. 1 by restricting "use" to students' individual use (individual participation) and by splitting "teachers' supply" into several areas: the task quality (held constant in this paper), the planned intended teacher activation, and interactional supply in teachers' enacted activation. Key to our research design was keeping the task quality and intended activation constant by supplying all groups with the same tasks, representations, and manipulatives and the same planned teacher moves (presented in the next subsection). We conceptualize interaction quality with eight subdimensions and operationalize them into 14 quality features (to be introduced below). We study how they are connected to students' learning gains on fractions, while distinguishing the potential relevance of different class contexts, as suggested in the literature (Bostic et al., 2021; Cai et al., 2020; Pauli & Reusser, 2015). All components are further explained in the next subsections.

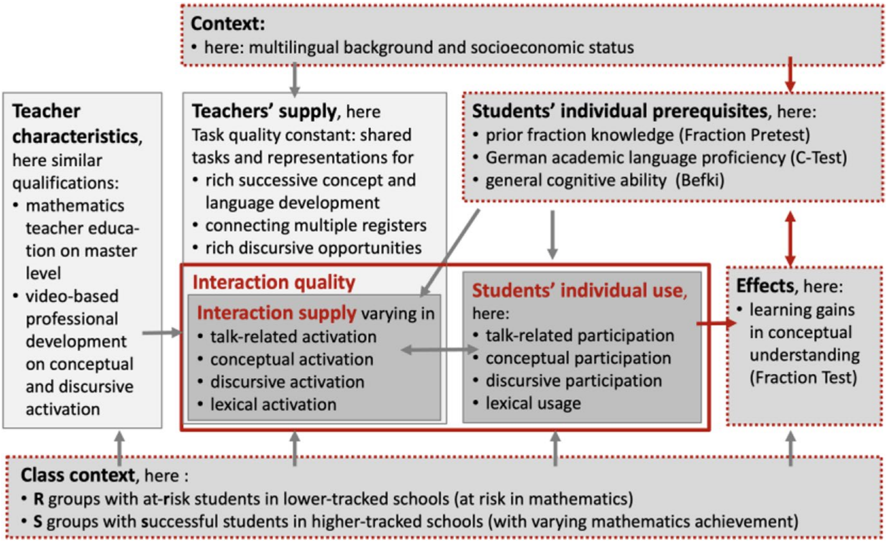
Context:
• here: multilingual background and socioeconomic status
Teacher characteristics, here similar qualifications:
• mathematics teacher education on master level
• video-based professional development on conceptual and discursive activation
Teachers' supply, here
Task quality constant: shared tasks and representations for
• rich successive concept and language development
• connecting multiple registers
• rich discursive opportunities
Students' individual prerequisites, here:
• prior fraction knowledge (Fraction Pretest)
• German academic language proficiency (C-Test)
• general cognitive ability (Befki)
Interaction quality
Interaction supply varying in
• talk-related activation
• conceptual activation
• discursive activation
• lexical activation
Students' individual use, here:
• talk-related participation
• conceptual participation
• discursive participation
• lexical usage
Effects, here:
• learning gains in conceptual understanding (Fraction Test)
Class context, here :
• R groups with at-risk students in lower-tracked schools (at risk in mathematics)
• S groups with successful students in higher-tracked schools (with varying mathematics achievement)

Fig. 1 Research design illustrated in the framework of a supply-use model adapted for this study's specific context of understanding of fractions

Constant Intended Task Quality in a Small Group Fraction Intervention

The data corpus used in this study originates from the intervention study MuM-MESUT, with a cognitively demanding and instructionally supportive intervention aiming at developing conceptual understanding of fractions and their operations and aiming at developing a bridging language for explaining meanings (Prediger et al., 2022). Instructional support and conceptual richness are provided in the tasks by a carefully designed conceptual learning trajectory and by connecting multiple representations (fraction bars for the part-whole concept and for exploring equivalent fractions, manipulatives for fractions of sets, etc., being connected to meaning-related language, context problems, and symbolic expressions; see Cramer et al., 1997). Cognitive demands were also distinguished by discursive richness, i.e. by tasks that systematically engage students in rich discourse practices and support their language development by the use of language-responsive design principles (Erath et al., 2021). 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 whole-class teaching (Prediger et al., 2022).

In order to capture each student's individual participation in detail, the instruction was organized in teacher-led small groups (3–6 students) and spanned over five videotaped sessions of 90 min each, taught by master's and PhD students with mathematics teaching certificates. By keeping the tasks and representations and suggested prompts nearly identical, which also kept the teachers' intended activations mostly the same, differences in teachers' enacted activations in the interactions and students' individual participation became observable in a fine-grained way, which also shaped the enacted task-based interaction quality.

Measures for Fraction Knowledge and Students' Individual Prerequisites

The following quantitative measures were administered prior to or after the intervention:

- *Fraction knowledge in pretest and posttest.* Students' conceptual understanding of fractions (the dependent variable) was measured by a standardized fraction test, covering the addressed aspects of conceptual understanding of fractions as well as procedures with fractions beyond the intervention content. The pretest (conducted before the intervention) and posttest (conducted after the intervention) contained the same items with different numbers. They had satisfactory internal consistencies of Cronbach's $\alpha = 0.82$ for the pretest with 25 items and a maximum score of 25 in our initial sample ($N = 1{,}403$) and $\alpha = 0.83$ for the posttest ($N = 721$).
- *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 (Grotjahn et al., 2002). It had a satisfactory internal consistency of Cronbach's $\alpha = 0.788$ ($N = 1403$). The maximum number of correctly filled gaps was 60.

- *General cognitive ability.* Fluid intelligence was measured using a matrix test (BEFKI 7). This test had a satisfactory internal consistency of Cronbach's $\alpha = 0.78$ (16 items, $N = 1403$).
- *Multilingual background.* Multilingual students were those who reported speaking multiple family languages or a family language other than the language of instruction.
- *Socioeconomic status.* Students' SES (known as a relevant factor for achievement; Reiss et al., 2019) was measured using the book-at-home index levels, asking students how many books they had at home with example photos (re-test reliability of $r = 0.80$, Levels 1–5).

Sample and Sampling

As Cai et al. (2020) called for conducting instructional quality research with a higher sensitivity to contexts, we considered different school contexts in our sample that reflected the selective German early tracking system, in which from Grade 5 (around 10 years old) onward, 40% of students are enrolled in higher-tracked schools (*Gymnasium*) and 60% in lower-tracked schools. Being enrolled in a higher-tracked school is a privileging factor being shown to create a school context with favorable learning opportunities (Reiss et al., 2019). In contrast, not all students in lower-tracked schools are at risk of being left behind in mathematics, so we focused only on those students in lower-tracked schools who were underachieving in mathematics.

In our sampling procedure, we included both school contexts in our initial sample of the overarching intervention study (Prediger et al., 2022): first, a subsample $\mathbf{R'}$ of seventh-graders at risk ($n = 323$) was selected among the weak mathematics achievers from lower-tracked schools (with a fraction pretest score below 15). The subsample $\mathbf{S'}$ ($n = 266$) of successful students was selected among 279 students from higher-tracked schools (academic success being operationalized by the higher track). In order to find enough students with a fraction pretest score below 15, we selected sixth graders before their systematic exposure to fractions.

In the second step of our sampling procedure, students in the samples $\mathbf{R'} + \mathbf{S'}$ were assigned to intervention and control groups in a cluster-randomized way. Of the intervention group with 394 students (in 92 small groups of 3 to 6 students each), 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 ($n = 210$) in these 49 groups formed our video sample for this paper. The descriptive characteristics of the resulting video sample with its subsamples $\mathbf{R}$ (students at risk) and $\mathbf{S}$ (successful students) are listed in Table 2, as they bear different characteristics, the analysis will also take into account potential interaction effects.

Table 2 Descriptive data for the video sample

Variable M (SD) or percent	Full video sample ($n=210$)	…in at-risk school contexts **R** ($n=83$)	… in successful school contexts **S** ($n=127$)
Fraction pretest score	7.64 (3.8)	8.88 (3.17)	6.81 (3.97)
Fraction posttest score	13.35 (4.28)	12.46 (4.2)	13.96 (4.25)
General cognitive ability	8.82 (3.8)	8.27 (2.74)	9.19 (4.34)
Academic language proficiency	40.04 (9.16)	37.19 (8.53)	41.94 (9.11)
Age	11.53 (1.15)	12.79 (0.62)	10.71 (0.49)
Multilingual Background (in %)	49	52	48
SES: low/medium/high (in %)	21/32/47	36/36/28	11/30/60

Quality Subdimensions and Quality Features with Different Operationalizations

Summing up, the video data corpus analyzed for this study stemmed from 49 teacher-led small groups (of 3–6 students each) who were all taught with the same curriculum material and the same teacher preparation. The shared tasks and representations ensured a constant intended task quality, and the shared teacher preparation ensured comparability in teachers' intended activations (first column of Table 1). This allowed us to scrutinize the impact of the considerable differences found in teachers' enacted activations and students' individual participation (second and third column) for the four quality domains represented in the rows of Table 1.

In this way, we could focus our study on eight subdimensions of cognitive demands and instructional support: 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*).

Basic Ratings for Richness on the Level of Tasks, Moves, and Practices

As the existing coding protocols draw upon different operationalizations or holistic ratings that have been criticized due to loss of information (Ing & Webb, 2012; Pauli & Reusser, 2015), we increased transparency and systematically compared different operationalizations for each of the quality subdimensions. The talk-related dimensions were operationalized as a baseline, independent of the richness of tasks, moves, and practices (as, for example, in Sedova et al., 2019) by the relative length of students' utterances: *talk-related participation* (*TP*) of individual students was measured by their relative individual talk time as percentage of the time on task and *talk-related activation* (*TA*) was measured by class engagement, in other words, the sum of all students' relative talk times as percentage of time on task. For all other subdimensions, the richness of the talk must be taken into account, operationalized by the richness of tasks, teacher moves, or students' and teachers' co-constructed practices. Basic (low and high inferent) ratings were developed for conceptual, discursive, and lexical richness on three levels:

- *Basic ratings for richness of tasks.* The *tasks* were rated regarding their *conceptual richness* (conceptually poor tasks focus on facts and procedures and conceptually rich tasks focus on conceptual understanding; rating followed Kunter et al., 2013), *discursive richness* (discursively poor tasks demand short half-sentence or one-word answers and discursively rich tasks demand students to explain meanings or report procedures; rating followed Wessel and Erath, 2018), and *lexical richness* (lexically rich tasks explicitly promote lexical learning, e.g. by asking students to reflect, collect, or use key phrases; following criteria of Carlisle et al., 2013). The design of the intervention included decisions about the tasks' intended conceptual, discursive, and lexical richness, thus the rating is low inferent. Based on this basic rating, *task-based quality features for the interaction* were derived by capturing the length of interaction time spent on tasks of a particular degree of richness. An example is the quality feature *CA-t* (abbreviation for *C*onceptual *A*ctivation in *t*ask-based operationalization; see second column of Table 3): the relative length of the group's time spent on conceptually rich tasks (instead of procedural tasks).
- *Basic ratings for richness of moves.* Teachers' *moves* were rated regarding their *conceptual richness* (conceptually poor moves ask and strengthen the focus on facts and procedures and conceptually rich moves ask for, support, or strengthen aspects of conceptual understanding) and *lexical richness* (lexically rich moves explicitly promote lexical learning, e.g. by asking students to reflect, collect, or use key phrases or by connecting or revoicing students' utterances with new vocabulary; following criteria of Carlisle et al., 2013; Kohlmeier, 2018). Based on these highly inferent 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. For example, *LA-m* entails the relative group time in which the teacher repeatedly engages the students utilizing moves supporting lexical learning out of the time on task, for instance, by demanding the meaning-related language for the part-whole concept.
- *Basic ratings for richness of practices.* All teachers' and students' utterances were rated highly inferent with respect to the richness of the collectively established discourse *practices* they contributed to regarding *discursive richness* (discursively rich oral discourse practices engage students in longer utterances elaborating an idea or reporting a procedure, whereas discursively poor discourse practices are constrained to brief contributions of half a sentence) and *conceptual richness* (conceptually rich discourse practices are those discursively rich discourse practices in which students explain meanings or describe mathematical structures; see Wessel and Erath, 2018). 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. One example is *DA-p* (last column in Table 3), which is the relative length of group talk spent on rich discourse practices, for example, negotiating mathematical meanings, out of the total time on task.

Table 3 Framework for features for quality interaction in task-, move-, and practice-based operationalizations (Relative length of talk time spent in each aspect of richness, given in percent out of total time on task)

	Quality features in task-based operationalization -t	**Quality features in move-based operationalization -m**	**Quality features in practice-based operationalization -p**
Talk-related activation	*TA* Relative length of all students' talk (with a varying richness of tasks, moves, and practices)		
Talk-related participation	*TP* Relative length of individual talk (with a varying richness of tasks, moves, and practices)		
Conceptual activation	*CA-t* Relative length of group time spent on conceptually rich tasks	*CA-m* Relative length of group time spent on conceptually rich moves	*CA-p* Relative length of group talk spent on conceptually rich practices
Conceptual participation	*CP-t* Relative length of individual talk spent on conceptually rich tasks		*CP-p* Relative length of individual talk spent on conceptually rich practices
Discursive activation	*DA-t* Relative length of group time spent on discursively rich tasks		*DA-p* Relative length of group talk spent on rich discourse practices
Discursive participation	*DP-t* Relative length of individual talk spent on discursively rich tasks		*DP-p* Relative length of individual talk spent on rich discourse practices
Lexical activation	*LA-t* Relative length of group time spent on tasks for lexical learning	*LA-m* Relative length of group time spent on moves sup-porting lexical learning	
Lexical participation	*LP-t* Relative length of individual talk spent on tasks for lexical learning		

All highly inferent basic ratings (of conceptual and lexical moves and discourse practices in 30 h of video data) were coded in the analysis software Transana independently by two raters (well-trained students on their way to master's degrees in mathematics education). Interrater reliability was controlled in R Studio (version 3.6.3, package DescTools), and the determined Cohen's κ of between 0.80 and 0.91 indicated that interrater reliability was very good (Döring & Bortz, 2016).

Rating Utterances According to Richness of Tasks, Moves, and Practices

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. As many students' contributions first deviated from the conceptual, discursive, and lexical richness demanded in teachers' moves, the individual participations were not operationalized in move-based ways.

Quality Features Measured by the Time Spent in Differently Rich Utterances

Existing coding protocols have used heterogeneous measurements for summarizing participants' utterances, in turn-related, sentence-related, and time-related frequencies; this heterogeneity limits the comparability within and across studies (Pauli & Reusser, 2015). To provide a unified measurement, all quality features were measured by time-related relative frequencies, which means by the percent of talk time spent for a certain degree of richness (e.g. a conceptual teacher move) in relation to the total time on task (including times of writing or silence).

With these basic ratings and decisions on unified time-related measurements, the eight conceptualized subdimensions 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, as listed in Table 3.

Methods for the Data Analysis

To analyze how the quality features had an impact on students' learning gains when all teachers shared the same tasks and representations, we tested whether students clustering in groups made multilevel analysis necessary. The intraclass correlation of $\mathrm{ICC}=0.12$ for the dependent variable of students' fraction posttest scores can be considered small in the context of small student groups (Hox, 2010). This allowed us to decide against a multilevel approach due to our small sample size, which would not have made the analysis of cross-level interaction effects in multilevel modeling possible (Hox, 2010). Instead, we conducted hierarchical multiple regression models with students' posttest scores as the dependent variable. To test the robustness of our results, we reanalyzed our data and the same effects of the quality features as with our multiple regression models (except we could not test the interaction effects).

In a pre-analysis, we tested with regression analysis how much students' fraction pretest score, their school context, their multilingual background, their SES, their general cognitive ability, and their academic language proficiency could predict their fraction posttest scores. We then built hierarchical regression models: Step 1 introduced the significant individual prerequisites in one basic regression model and Step 2 added one quality feature and its interaction effect with the school context each. The 14 quality features (and interaction terms) were entered separately into 14 models, as they were highly correlated (Quabeck et al., 2023).

All metric independent variables were z-standardized. Variables were checked for outliers and unusual cases. In one case, Cook distances for the regression models were close to the cut-off of 1, so we re-ran all analyses without this case and could not observe any substantial changes in model fits or estimated parameters (Field, 2013). Therefore, we kept this case in our data set. Assumptions for conducting regression analysis were checked with graphs and tests for each model (linearity, independent errors, normally distributed errors, homoscedasticity, and multicollinearity; Field, 2013). We tested each model to determine whether it explained variance significantly and also whether the added quality dimensions and their interaction with the school context added significantly to the explained variance in students' posttest scores. Standardized beta coefficients β were also determined for each independent variable.

Results

Distribution of 14 Enacted Quality Features for Interaction

The descriptive results of the 14 quality features in the 49 small groups are documented in Table 4. The means refer to the average time a group/a student spent

Table 4 Distribution of enacted quality features with mean (and SD) for relative lengths in the 49 groups

Quality feature	**Operationalization of the quality features** (Relative length of… in relation to total time on task of the group)		**M (SD)** of relative length
Talk-related activation	TA	All students' talk	33.4% (14%)
Talk-related participation	TP	Individual talk	7.7% (5%)
Conceptual activation	CA-t	Group time spent on conceptually rich tasks	77.4% (11%)
	CA-m	Group time spent on conceptually rich moves	35.3% (13%)
	CA-p	Group talk on conceptual practices	23.5% (12%)
Conceptual participation	CP-t	Individual talk spent on conceptually rich tasks	5.5% (4%)
	CP-p	Individual talk spent on conceptual practices	2.8% (3%)
Discursive activation	DA-t	Group time spent on discursively rich tasks	11.0% (8%)
	DA-p	Group talk on rich discourse practices	33.0% (12%)
Discursive participation	DP-t	Individual talk spent on oral discursively rich tasks	1.3% (1%)
	DP-p	Individual talk spent on rich discourse practices	3.9% (3%)
Lexical activation	LA-t	Group time spent on tasks for lexical learning	10.7% (9%)
	LA-m	Group time spent on moves on lexical learning	42.4% (13%)
Lexical participation	LP-t	Individual talk spent on tasks with lexical learning opportunities	0.9% (1%)

on tasks/moves/practices qualified as rich (not qualified/conceptually rich/discursively rich); for instance, $M(CA\text{-}t) = 77.4\%$ means that on average, the groups spent 77.4% of their time on task on conceptually rich tasks. $SD(CA\text{-}t) = 11\%$ means that a typical dispersion in time was between 66 and 88%.

The data revealed that within the same subdimension, the 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* for 77.4% of the time) was more than the double of time spent on teachers' conceptual moves (*CA-m* for 35.3% of the time) or in established conceptual practices (*CA-p* for 23.5%). This means that even for tasks with conceptual focus, this focus was not always reflected in teachers' moves, and in teachers' and students' interactively co-constructed conceptual practices only in a third of the time. In contrast, although the time spent on tasks with explicit rich discursive task demands (*DA-t*) was only 11%, the group talk was engaged in rich discourse practices for 33% of the time (*DA-p*). Similarly, the teachers' moves strengthened the lexical activation: only 10.7% of the time was dedicated to tasks that explicitly provide lexical learning opportunities (*LA-t*), but 42.4% of the time, the groups were lexically activated by the teachers' lexically focused moves (*LA-m*).

Regression Models Predicting Student Learning in 14 Quality Features for Cognitively Demanding and Instructionally Supportive Interaction

In a first pre-analysis, all individual prerequisites were entered into one regression model for identifying predictors for students' scores in the fraction posttest. Students' fraction pretest score, their general cognitive ability, and the school context were significant predictors, while multilingual background, immigrant status, and academic language proficiency were not. The first three variables were used for the first step of our hierarchical models. In 14 further models, one quality feature and its interaction with the school context were added to the model. Assumptions for conducting multiple regression analyses were satisfied for every model.

Effects of Eight Quality Features About Teachers' Enacted Activation

The regression models for eight quality features of teachers' enacted activation are given in Table 5. The move-based measure for lexical activation (*LA-m*) was a significant positive predictor of students' posttest scores (Model 2h). Their $\beta = 0.19$ indicated that as lexical activation moves increased by one standard deviation, students' posttest scores increased by 19% of a standard deviation. Furthermore, the interaction between conceptual activation practices (*CA-p*) and at-risk students was predictive of student learning (Model 2d): while more time spent on these practices had a positive effect for at-risk students, it seems to have had a negative effect for successful students. No task-related quality feature significantly predicted the posttest scores. The regression models explain between 39% and 42% of the variance in posttest scores.

Table 5 Predictors in teachers' enacted activation for student learning: multiple regression models predicting fraction posttest scores from individual prerequisites and eight quality features on enacted activation

	β	F	R^2	Adj. R^2	ΔR^2
Model 1: base model		42.50***	.39	.38	-
Fraction pretest	0.53***				
General cognitive ability	0.18**				
At-risk context	– 0.30***				
Model 2a: talk-related activation		25.25***	.39	.38	.00
Fraction pretest	0.54***				
General cognitive ability	0.18**				
At-risk context	– 0.30**				
Talk-related activation (*TA*)	– 0.01				
Talk-related activation x at-risk context	0.01				
Model 2b: conceptual activation (task based)		25.26***	.39	.37	.00
Fraction pretest	0.53***				
General cognitive ability	0.18**				
At-risk context	– 0.30***				
Conceptual activation (task based) (*CA-t*)	– 0.01				
Conceptual activation (task based) x at-risk context	0.00				
Model 2c: conceptual activation (moved based)		25.95***	.40	.38	.01
Fraction pretest	0.54***				
General cognitive ability	0.18**				
At-risk context	– 0.28***				
Conceptual activation (move based) (*CA-m*)	0.11				
Conceptual activation (move based) x at-risk context	– 0.05				
Model 2d: conceptual activation (practice based)		28.523***	.42	.40	.03*
Fraction pretest	0.52***				
General cognitive ability	0.16**				
At-risk context	– 0.28***				
Conceptual activation (practice based) (*CA-p*)	– 0.13				
Conceptual activation (practice based) x at-risk context	0.21**				
Model 2e: discursive activation (task based)		26.31***	.40	.39	.01
Fraction pretest	0.54***				
General cognitive ability	0.19**				
At-risk context	– 0.28***				
Discursive activation (task based) (*DA-t*)	– 0.06				
Discursive activation (task based) x at-risk context	0.12				
Model 2f: discursive activation (practice based)		26.41***	.40	.39	.01
Fraction pretest	0.53***				
General cognitive ability	0.18**				

Table 5 (continued)

	β	*F*	R^2	Adj. R^2	ΔR^2
At-risk context	−0.28***				
Discursive activation (practice based) (*DA-p*)	−0.08				
Discursive activation (practice based) x at-risk context	0.14				
Model 2g: lexical activation (task based)		25.99***	.40	.38	.01
Fraction pretest	0.53***				
General cognitive ability	0.20**				
At-risk context	−0.29***				
Lexical activation (task based) (*LA-t*)	−0.01				
Lexical activation (task based) x at-risk context	0.09				
Model 2h: lexical activation (move based)		27.99***	.41	.40	.03*
Fraction pretest	0.55***				
General cognitive ability	0.16**				
At-risk context	−0.29***				
Lexical activation (move based) (*LA-m*)	0.19**				
Lexical activation (move based) x at-risk context	−0.09				

(*: $p<0.05$, **: $p<0.01$, ***: $p<0.001$)

Effects of Six Quality Features of Students' Individual Participation

The regression models for the six quality features of students' individual participation are given in Table 6. None of the six quality features predicted student learning when we controlled for students' pretest scores, their cognitive abilities, and their school context.

Conclusions

Discussion of the Findings

As various studies have revealed that cognitively demanding and supportive instruction can promote students' learning (Howe et al., 2019; Lipowsky et al., 2009), it is worth disentangling the existing heterogeneous conceptualizations and operationalizations of these quality dimensions (Cai et al., 2020; Praetorius & Charalambous, 2018). Our video data corpus confirmed that the quality cannot be captured by task quality alone (Henningsen & Stein, 1997; Schoenfeld, 2014), as the 49 small groups worked with the same tasks and representations of high task quality, but showed large standard deviations in the task-based measures of interaction quality (Table 4).

By the transparent disentanglement of four potentially relevant subdimensions of cognitively demanding and supportive interaction (Table 3), we contribute also to the methodological discourse on how to operationalize judgments of richness

Table 6 Predictors in students' individual participations for student learning: Multiple regression models predicting fraction posttest scores from individual prerequisites and six quality features on participation

	β	F	R^2	Adj. R^2	ΔR^2
Model 1: base model		42.50***	.39	.38	-
Fraction pretest	0.53***				
General cognitive ability	0.18**				
At-risk context	– .030***				
Model 2i: talk-related participation		25.44***	.39	.38	.00
Fraction pretest	0.53***				
General cognitive ability	0.18**				
At-risk context	– 0.30***				
Talk-related participation (*TP*)	0.05				
Talk-related participation x at-risk context	– 0.05				
Model 2j: conceptual participation (task-based)		25.61***	.39	.38	.00
Fraction pretest	0.53***				
General cognitive ability	0.19**				
At-risk context	– 0.29***				
Conceptual participation (task based) (*CP-t*)	0.07				
Conceptual participation (task based) x at-risk context	– 0.07				
Model 2k: conceptual participation (practice based)		25.66***	.39	.38	.00
Fraction pretest	0.53***				
General cognitive ability	0.19**				
At-risk context	– 0.28***				
Conceptual participation (practice based) (*CP-p*)	0.03				
Conceptual part. (practice based) x at-risk context	0.04				
Model 2l: discursive participation (task based)		28.523***	.40	.38	.01
Fraction pretest	0.53***				
General cognitive ability	0.17**				
At-risk context	– 0.27***				
Discursive participation (task based) (*DP-t*)	0.08				
Discursive participation (task based) x at-risk context	0.04**				
Model 2m: discursive participation (practice based)		26.31***	.39	.38	.00
Fraction pretest	0.53***				
General cognitive ability	0.18**				
At-risk context	– 0.29***				
Discursive participation (practice based) (*DP-p*)	0.08				
Discursive participation (practice based) x at-risk context	– 0.05				

Table 6 (continued)

	β	F	R^2	Adj. R^2	ΔR^2
Model 2n: lexical participation (task based)		25.33***	.39	.38	.00
Fraction pretest	0.54***				
General cognitive ability	0.18**				
At-risk context	−0.31***				
Lexical participation (task based) (*LP-t*)	0.00				
Lexical participation (task based) x at-risk context	−0.03				

(*: $p<0.05$, **: $p<0.01$, ***: $p<0.001$)

(Praetorius & Charalambous, 2018). The varying values in Table 4 indicate that they indeed measure different phenomena.

The research question asks for the effects of quality features on students' posttest scores when controlling for relevant prerequisites (fraction pretest scores and general cognitive ability). Given that the intervention as a whole with its high task quality already showed overall effectiveness for students' average learning gains (in a cluster-randomized trial in Prediger et al., 2022), we did not expect all quality features to reveal additional effects.

Yet it was unexpected that only 2 out of 14 quality features had an additional effect on students' posttest scores, move-based lexical activation (*LA-m*) for all students, and practice-based conceptual activation (*CA-p*) for students in at-risk contexts, which needs to be further discussed: None of the quality features of *students' individual participation* had an additional effect on students' learning gains (Table 6). It does not seem to matter how long the individual students speak (talk-related participation, *TP*), this is in line with existing qualitative findings (Ing & Webb, 2012; Walshaw & Anthony, 2008). It did also not matter how long the student spoke in the group talk about conceptually rich tasks (*CP-t*), discursively rich tasks (*DP-t*), or lexically rich tasks (*LP-t*), and not even in rich conceptual practices (*CP-p*) or rich discourse practices (*DP-p*). While this might at first be astonishing, it also resonates with qualitative findings that active individual participation does not necessarily mean talking, but can also include more silent forms of participation, for instance, as a listening discussion partner (Krummheuer, 2011). It is also in line with the socially co-constructed nature of practices in which individual contribution is less important than group collective achievement. This might be very different in small groups of two to six students than in whole classes with 30 students in which those who do not talk also have a higher risk of not following mentally.

The *enacted activation* was operationalized by the relative length of time on more or less *rich group talk*, with this richness being measured by the richness of the discussed tasks, the teachers' moves, and the interactively established practices. Among the eight quality features, the pure space to talk (*TA*, as in Inagaki et al., 1998) and those with task-based operationalizations of richness (*CA-t, DA-t, LA-t*) were not predictive for students' posttest scores (Table 5). The time spent on high-quality tasks does not impact the learning gains, as time does not grant high conceptual/discursive/lexical demands being maintained (Henningsen & Stein, 1997).

However, the instructional support provided by teachers' lexical moves seemed to matter (move-based lexical activation *LA-m*): the relative length of group time spent on *moves supporting lexical learning* significantly predicted the posttest scores (β = 0.19) after controlling for individual prerequisites and school context. *Lexical support* in students' language production has been promoted in contexts of language-responsive classrooms (Gibbons, 2002) and identified as a part of an instructional quality feature (Hill et al., 2008). On the level of task quality, interventions with and without explicit lexical support have been shown to be effective for students' learning gains in a previous paper (Prediger et al., 2022), which corresponds to the non-significance of the relative time spent in tasks with lexical activation (*LA-t*). But the predictive power of the quality feature *LA-m* means that the longer the small group engages after teachers' moves with lexical learning opportunities (e.g. using offered phrases in an argument or after teachers' revoicing of a student utterance in the target phrases), the more they learn. This finding resonates with earlier findings on the role of enhancing language production for deepening mathematics understanding (e.g. Gibbons, 2002; Smit et al., 2013), as the lexical work here was always discursively embedded.

Particularly for the at-risk students, the interaction term with conceptual practices (*R x CA-p* with ß = 0.21) was significantly predictive for the posttest scores. That means, the longer at-risk students are engaged in conceptual practices (e.g. in explaining meanings for mathematical concepts), the more they learn mathematically. The quality feature practice-based conceptual activation was not predictive for successful students (who might be able to accomplish a complete explanation quicker), and it was not true for all rich discourse practices (*DA-p*) that also involved, for instance, the reporting of procedures.

This finding supports for the often-identified necessary conceptual focus of teaching (Hiebert & Grouws, 2007; Schoenfeld, 2014): *CA-t*, the relative length of time spent on conceptual tasks (on average 77.4%; see Table 4) was not predictive for student learning, but *CA-p*, the relative length of time spent with conceptual practices (on average 23.5%), was predictive. Beyond this replication, the latter is a very strong support for the design principle of *engaging students in conceptually rich discourse practices* that has been promoted by many design researchers in the context of language-responsive mathematics teaching (Erath et al., 2021; Gibbons, 2002). The delineation between *CA-p* and *DA-p* also confirms the need to distinguish rich discourse practices referring to procedures from those referring to conceptual aspects (reporting procedures is less supportive for learning than explaining meanings). This distinction resonates with qualitative findings from earlier case studies (Moschkovich, 2015), but this paper is the first to also provide quantitative evidence for its relevance.

Methodological Limitations and Future Research Needs

Of course, the findings must be interpreted with respect to the methodological limitations. The first limitation is the given data set itself. With only 49 small groups of three to six students each, the sample size was not large enough to conduct multilevel models with interaction effects (as e.g. by Decristan et al., 2015). However, the multilevel models revealed the same non-significance of most quality features, with only *LA-m* being predictive for the posttest scores. In future studies, larger group numbers and larger group sizes should be investigated in order to allow for multilevel modelling and also for approaching more ecological validity of the interaction taking place in whole classes with 25–30 rather than groups with three to six students (as Howe et al., 2019, requested).

Another limitation might be that the time-related measurement of the activation and participation cannot sufficiently reflect the relevance of some moments over others. Even if the time-related measurement of talk time qualified by conceptual, discursive, or lexical richness already requires a huge amount of coding time, it remains a rough approximation for the qualities identified in qualitative studies (Walshaw & Anthony, 2008). Future studies could explore other task-based, move-based, or practice-based measurements and also overcome gaps in our table of investigated operationalizations (with missing *LA-p*, *LP-m*, *DA-m*, and *DA-m*). Similarly, active participation might not always be visible. Students might be engaged in learning even though they are not actively (e.g. verbally) participating but learning by observing others (Krummheuer, 2011). These limitations might be tackled by future face-scanning technologies.

Furthermore, the operationalizations of interaction quality do not capture adaptability or tell us why more or less time was spent (some students might spend more time talking about conceptual aspects as they struggle with a concept, while others might explain a concept comprehensively), but this can only be captured in qualitative analysis. Finally, the study should be extended to other teaching materials, other school contexts, and to other mathematical topics in order to investigate their transferability more systematically than has been done so far.

Implications for Teacher Professional Development and Future Research

For future research on interaction quality, we hope that our in-depth scrutinizing of subdimensions into quality features with transparent operationalizations will also motivate other researchers to dive deeper into the details of capturing high-quality interaction quantitatively. Only after many further studies in other contexts and with other teachers and other mathematical topics will consolidated findings accumulate over time.

The current findings can already have substantial consequences for teachers' professional development in foregrounding interaction quality, for example, discussing video excerpts with respect to conceptual, discursive, and lexical richness might be a promising pathway to disentangling the general ideas of cognitively demanding and supportive teaching (Schoenfeld, 2014). So far, many professional development programs seem to have focused mainly on talk-related dimensions, in other words, the quantitative space for student talk, with an implicit assumption that the more students talk, the more chances they have for rich talk. However, even within a narrow intervention program, the relative length of talk qualified as conceptually rich, discursively rich, and lexically rich was so widespread that teachers should be made aware of these differences. The fact that the *lexical activation following teachers' moves* is the most predictive quality feature needs to be treated in professional development on language-responsive mathematics teaching to enable teachers to provide the lexical support in their moves, which turned out to be more crucial than the tasks or simple vocabulary lists (Prediger et al., 2022). Finally, *activating conceptual practices of collectively explaining meanings and arguing about connections* should play a major role in professional development, as this is highly important, particularly for students at risk.

Funding Open Access funding enabled and organized by Projekt DEAL. The project MESUT 2 (Developing conceptual understanding by language support: Studying differential conditions of success in the supply-use-model) was funded by the German Research Foundation under DFG-grants PR662/14–2 to S. Prediger and ER880/3–3 to K. Erath.

Data Availability Data is available in German language upon request from the authors.

Declarations

Competing Interests The authors declare no competing interests.

References

Allen, J., Gregory, A., Mikami, A., Lun, J., Hamre, B., & Pianta, R. (2013). Observations of effective teacher-student interactions in secondary school classrooms. *School Psychology Review, 42*(1), 76–98. https://doi.org/10.1080/02796015.2013.12087492

Bauersfeld, H. (1988). Interaction, construction, and knowledge: Alternative perspectives for mathematics education. In D. A. Grouws & T. J. Cooney (Eds.), *Perspectives on research on effective mathematics teaching: Research agenda for mathematics education* (pp. 27–46). NCTM/Lawrence Erlbaum.

Bostic, J., Lesseig, K., Sherman, M., & Boston, M. (2021). Classroom observation and mathematics education research. *Journal of Mathematics Teacher Education, 24*(1), 5–31. https://doi.org/10.1007/s10857-019-09445-0

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

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

Carlisle, J. F., Kelcey, B., & Berebitsky, D. (2013). Teachers' support of students' vocabulary learning during literacy instruction in high poverty elementary schools. *American Educational Research Journal, 50*(6), 1360–1391. https://doi.org/10.3102/0002831213492844

Cramer, K., Behr, M., Post, T., & Lesh, R. (1997). *Rational number project: Fraction lessons for the middle grades.* Kendall/Hunt.

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

Döring, N., & Bortz, J. (2016). *Forschungsmethoden und evaluation in den Sozial- und Humanwissenschaften* [Research methods and evaluation in the social and human sciences]. Springer. https://doi.org/10.1007/978-3-642-41089-5

Erath, K., Ingram, J., Moschkovich, J. N., & Prediger, S. (2021). Designing and enacting instruction that enhances language for mathematics learning. *ZDM – Mathematics Education, 53*(2), 317–335. https://doi.org/10.1007/s11858-020-01213-2

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.

Grotjahn, R., Klein-Braley, C., & Raatz, U. (2002). C-test: An overview. In J. A. Coleman, R. Grotjahn, & U. Raatz (Eds.), *University language testing and the C-test* (pp. 93–114). AKS Finkenstaedt.

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

Henningsen, M., & Stein, M. K. (1997). Mathematical tasks and student cognition. *Journal for Research in Mathematics Education, 28*(5), 524–549. https://doi.org/10.2307/749690

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.

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., Hennessy, S., Mercer, N., Vrikki, M., & Wheatley, L. (2019). Teacher-student dialogue during classroom teaching: Does it really impact on student outcomes? *Journal of the Learning Sciences, 28*(4–5), 462–512. https://doi.org/10.1080/10508406.2019.1573730

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications*. Routledge.

Hsu, H.-Y., Yao, C.-Y., & Lu, B. (2023). Examination of Taiwanese mathematics teacher questioning. *International Journal of Science and Mathematics Education, 21*(5), 1473–1493. https://doi.org/10.1007/s10763-022-10313-2

Inagaki, K., Hatano, G., & Morita, E. (1998). Construction of mathematical knowledge through whole-class discussion. *Learning and Instruction, 8*(6), 503–526. https://doi.org/10.1016/S0959-4752(98)00032-2

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

Kohlmeier, T. L. (2018). *Instructional support for vocabulary acquisition among young dual language learners* [Doctoral dissertation]. Utah State University.

Krummheuer, G. (2011). Representation of the notion "learning-as-participation" in everyday situations of mathematics classes. *ZDM – Mathematics Education, 43*, 81–90. https://doi.org/10.1007/s11858-010-0294-1

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. https://doi.org/10.1007/978-1-4614-5149-5

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

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

Neugebauer, P., & Prediger, S. (2023). Quality of teaching practices for all students: Multilevel analysis of language-responsive teaching for robust understanding. *International Journal of Science and Mathematics Education, 21*(3), 811–834. https://doi.org/10.1007/s10763-022-10274-6

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

Organization for Economic Cooperation and Development [OECD]. (2020). *Global teaching insights: A video study of teaching*. OECD. https://doi.org/10.1787/20d6f36b-en

Pauli, C., & Lipowsky, F. (2007). Mitmachen oder zuhören? Mündliche Schülerinnen- und Schülerbeteiligung im Mathematikunterricht [Participate or listen? Oral student contributions in mathematics classrooms]. *Unterrichtswissenschaft, 35*(2), 101–124. https://doi.org/10.25656/01:5488

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* (pp. 181–193). 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

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

Reiss, K., Weis, M., Klieme, E., & Köller, O. (2019). *PISA 2018: Grundbildung im internationalen Vergleich* [PISA 2018: Basic education in international comparison]. Waxmann.

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

Sedova, K., Sedlacek, M., Svaricek, R., Majcik, M., Navratilova, J., DrexlerovaJ, 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

Smit, J., van Eerde, H. A. A., & Bakker, A. (2013). A conceptualisation of whole-class scaffolding. *British Educational Research Journal, 39*(5), 817–834. https://doi.org/10.1002/berj.3007

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

Stein, M. K., & Lane, S. (1996). Instructional tasks and the development of student capacity to think and reason. *Educational Research and Evaluation, 2*(1), 50–80. https://doi.org/10.1080/1380361960020103

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

Vieluf, S., Praetorius, A.-K., Rakoczy, K., Kleinknecht, M., & Pietsch, M. (2020). Angebots-Nutzungs-Modelle der Wirkweise des Unterrichts [Supply-use models for the effects of teaching]. *Zeitschrift für Pädagogik, 66*(Suppl.), 63–80. https://doi.org/10.3262/ZPB2001063

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

Wessel, L., & Erath, K. (2018). Theoretical frameworks for designing and analyzing language-responsive mathematics teaching-learning arrangements. *ZDM – Mathematics Education, 50*(6), 1053–1064. https://doi.org/10.1007/s11858-018-0980-y

Zhou, J., Bao, J., & He, R. (2023). Characteristics of good mathematics teaching in China: Findings from classroom observations. *International Journal of Science and Mathematics Education, 21*(4), 1177–1196. https://doi.org/10.1007/s10763-022-10291-5