International Journal of Science and Mathematics Education (2023) 21:811–834
https://doi.org/10.1007/s10763-022-10274-6

Quality of Teaching Practices for All Students: Multilevel Analysis of Language-Responsive Teaching for Robust Understanding

Philipp Neugebauer[1] · Susanne Prediger[1,2]

Received: 1 July 2021 / Accepted: 5 April 2022 / Published online: 30 April 2022

Abstract

The quality of curriculum resources and teaching practices can constrain or promote students' opportunities for mathematics learning, in particular, students with diverse language proficiency. The video study investigates 18 classes that all used the same curriculum resources aimed at developing 367 seventh graders' conceptual understanding of percentages to identify the interaction of quality dimensions, the enactment of given curriculum resources, and students' mathematical achievement (when controlling for mathematical preknowledge and language proficiency). Multilevel regression analysis revealed that three quality dimensions that can easily be supported by the curriculum resources (*Mathematical Richness*, *Cognitive Demand*, and *Connecting Registers*) were on a high level, and their variance had no additional interaction with students' achievement. In contrast, the 4 quality dimensions that were enacted in the teacher-student interaction with more variance (*Agency*, *Equitable Access*, *Discursive Demand*, and, in particular, *Use of Student Contributions*) had a significant additional impact on student achievement. These findings reveal important insights into the implementability of equitable instructional approaches.

Keywords Learning opportunities · Teaching practices · Language proficiency · Instructional quality

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

Philipp Neugebauer
philipp2.neugebauer@tu-dortmund.de

[1] TU Dortmund University, Vogelpothsweg 87, 44227 Dortmund, Germany

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

Introduction: Research Agenda of Investigating Quality Learning Opportunities

What students can achieve in mathematics classrooms is "ultimately determined and constrained by the opportunities they have had to learn" (Kilpatrick et al., 2001, p. 31). For this reason, a recent JRME editorial focused on one major problem in mathematics education research: "understanding how to maximize the quality of learning opportunities for every student" (Cai et al., 2020, p. 13). The editorial calls for research that helps to disentangle the quality dimensions of learning opportunities and their effects on students with diverse abilities. The urgency of this problem has also been underlined by repeated findings that students of different backgrounds have unequal access to higher quality opportunities (Callahan, 2005; Wilhelm et al., 2017), so maximizing the quality of learning opportunities for every student puts an additional emphasis on (a) ensuring that all students have access to quality opportunities and (b) the need to take into consideration students' diverse abilities, such as language proficiency and gender.

Going beyond existing research findings on the quality of instruction in general (e.g. Brophy, 2000; Charalambous & Praetorius, 2018; Hill et al., 2008), the JRME editorial calls for deepening the research, with a particular focus on various marginalized groups of students and with a more substantiated conceptualization of learning opportunities. The editorial suggests conceptualizing classroom-based learning opportunities as interactions between tasks, teaching, and students and following the overarching research interest in "what kinds of interactions among tasks, teaching, and students create learning opportunities for a specific learning goal" (Cai et al., 2020, p. 16). This paper contributes to this long-term research agenda, which is relevant not only to the USA but to many countries, including Germany, the country of the study in view. By investigating the interaction between teaching practices and students' learning in classrooms with the same curriculum resources, the impact of quality teaching practices in classrooms can be scrutinized.

The theory section of this paper outlines the state of research underlying the methodological choices in our research strategy and the analytic framework of language-responsive teaching for robust understanding (L-TRU). The method section presents the research design and the video rating. The empirical section presents the findings of the multilevel analysis of quality dimensions that have an impact on student learning.

State of Research and Theoretical Framework

Research on the Quality of Instruction and Learning Opportunities

Forty years of international research on the effectiveness of instruction in mathematics education have led to a consensus that surface structures such as activity structures (group work vs. seat work, etc.) are less relevant for the quality of learning opportunities than deeper structures involving, for example, the quality of tasks, adaptivity of teacher feedback, and suitability of representations used (Seidel & Shavelson, 2007). A survey on quality of instruction research (Charalambous & Praetorius, 2018) revealed that beyond the frameworks that capture generic instructional quality (e.g. the three basic dimensions of classroom management, student support, and cognitive activation; Praetorius et al., 2018), subject-related frameworks that involve subject-specific aspects such as mathematical richness and correctness, dealing with multiple representations, and appropriateness of examples have increasingly gained importance (e.g. Adler & Ronda, 2015; Decristan et al., 2015; Hill et al., 2008). Hiebert and Grouws (2007) in their survey emphasized the dimensions cognitive demand and focus on conceptual understanding. Some studies have assessed cognitive demands in classrooms by analyzing the tasks and curriculum resources used (Kunter et al., 2013), while others have also captured the ways cognitive demands are maintained in teacher-student interaction, as curriculum resources and teaching practices need not be completely aligned: whereas working with tasks with low cognitive demands rarely results in rich teaching practices, the potential of rich tasks is not necessarily exploited when enacted with unproductive teaching practices in the interaction (Stein et al., 2000). These findings suggest that instruction and learning opportunities should be considered an interplay of content, curriculum resources, teaching, and students (Cai et al., 2020; Cohen et al., 2003).

Students' Diverse Language Proficiencies and Learning Opportunities

Cai et al. (2020) emphasize that research on the quality of learning opportunities should involve perspectives on students' diverse abilities. One critical diversity factor is access to academic language as it has been shown to coincide with limited mathematics achievement, for both multilingual students (Barwell et al., 2016) and monolingual students from underserved communities (Moschkovich, 2010). In many countries, the language-related achievement gap can be traced back to opportunity gaps, as underserved communities (with many students with limited language access belong) often receive low-quality instruction, with an exclusive focus on procedural knowledge, low cognitive demands, and an incoherent curriculum (Callahan, 2005; Ing et al., 2015; Secada, 1992).

But even in mathematically rich, conceptually oriented classrooms with high cognitive demands, individual opportunity gaps can occur for students with limited academic language, when teachers fail to enable them to participate and exploit the learning opportunities provided (Herbel-Eisenmann et al., 2011). As a result, the conceptual understanding of mathematics of students with limited academic language often lags behind that of more language proficient peers, even more than it does for procedural skills (Moschkovich, 2015; Prediger et al., 2018). This requires instructional strategies aimed at enhancing language-related learning opportunities:

The survey on instruction that enhances language for mathematics learning summarizes language-responsive design principles for curriculum resources that can enable all students (nowithstanding their language proficiency) to exploit the provided learning opportunities (Erath et al., 2021). Qualitative and quantitative empirical evidence exists for the effectiveness of four main design principles: (1) engaging students in rich discourse practices and supporting their participation, (2) connecting language registers and multimodal representations, (3) using macro-scaffolding to sequence and combine language and mathematics learning opportunities, and (4) comparing and contrasting language aspects (form, function, etc.) to raise students' language awareness. These design principles for curriculum resources and interventions have proven effective in controlled trials, yet so far mainly without controlling for the teaching practices used when they are enacted. Another research tradition has investigated teaching practices, mainly in qualitative studies, and often in classrooms without teacher support from language-responsive curriculum resources (see surveys by de Araujo et al., 2018; Herbel-Eisenmann et al., 2011; Erath et al., 2021). They have identified productive practices for supporting students' equitable access to mathematics (e.g. using visuals, revoking, and leaving space for including cultural resources) and promoting students' agency.

To sum up, existing research on learning opportunities for students with limited academic language has often been separated into two areas: (a) design research and controlled trials based on design principles for language-responsive curriculum resources and (b) mainly qualitative observation studies on teaching practices (often conducted in classrooms with low-quality curriculum resources and without quantitative evidence). Overcoming this separation calls for studying teaching practices in classrooms with curriculum resources particularly optimized for enhancing language.

Research Strategy for Investigating the Quality of Learning Opportunities, Research Question, and the L-TRU Framework for Capturing Quality

The state of research on instructional quality for all students and with respect to language proficiency justifies three choices in our research strategy by which we intend to contribute to the research agenda suggested by Cai et al. (2020) for the chosen diversity factor of language proficiency:

(1) Cai et al. (2020) suggested starting with normative choices on the particular mathematical content goals in view. As the language-related opportunity gap has been shown to be bigger for conceptual understanding than for procedural skills, we focus on robust conceptual understanding of percentages, a topic crucial for middle school mathematics and mathematical literacy outside school but with challenges in understanding (Parker & Leinhardt, 1995).

(2) To investigate the interplay of content, curriculum resources, students, and teaching, we reduce the complexity by investigating the interaction of two components, teaching and student abilities, while keeping constant the content and the curriculum resources used for all observed classes. This allows us to focus on the impact of quality teaching practices in classrooms because Kilpatrick (2003) emphasized that "two classrooms in which the same curriculum is supposedly

being "implemented" may look very different; the activities of teacher and students in each room may be quite dissimilar, with different learning opportunities available, different mathematical ideas under consideration, and different outcomes achieved" (p. 473).

(3) We particularly focus on the diversity factor of language proficiency by optimizing the shared curriculum resources using language-responsive design principles (Erath et al., 2021) and by capturing the instructional quality using an analytic framework optimized for language-responsive teaching for robust understanding, which will be presented below.

By these choices in the research strategy, we refine the general research question on interactions among tasks, teaching, and students and learning opportunities (Cai et al., 2020, p. 16) as follows:

> What kinds of interactions between the quality of teaching practices and students' abilities (in language proficiency) create learning opportunities (for robust understanding of percentages) when the curriculum resources are held constant (with a focus on enhancing the mathematics learning of students with diverse language proficiency)?

Language-Responsive Teaching for Robust Understanding (L-TRU) Framework

To also capture the quality of teaching practices with respect to enhancing students' language for mathematics learning, we have adapted Schoenfeld's (2013) TRU framework, teaching for robust understanding, to include language as the L-TRU framework (Prediger & Neugebauer, 2021). Schoenfeld developed the TRU framework for rating quality dimensions of teaching practices. It was first developed for reflection in professional development, and it later also evolved into a more widely used analytic tool for research purposes (Schoenfeld, 2013; Schoenfeld et al., 2018). The TRU framework starts from *Mathematical Richness* (as Hill et al., 2008). From there, it unfolds students' experiences with mathematics in four more dimensions:

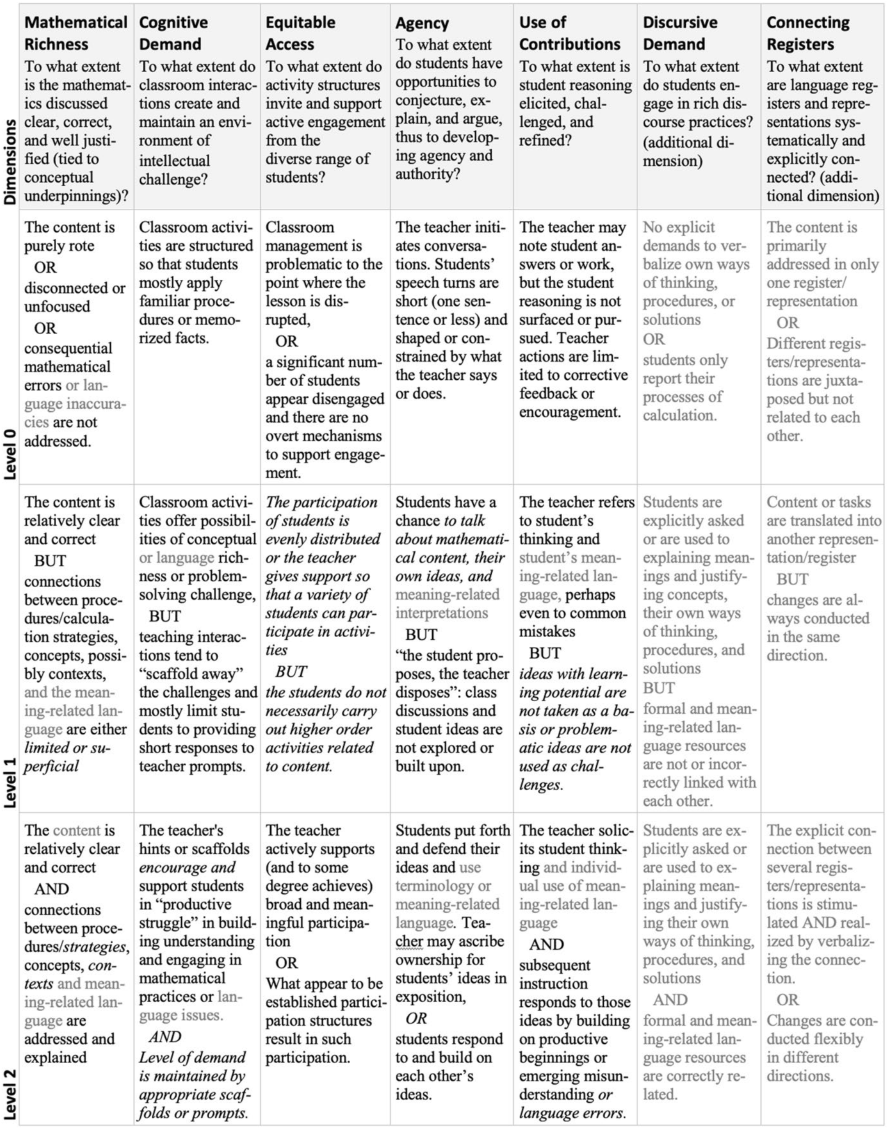
- *Mathematical Richness*: To what extent is the mathematics discussed clear, correct, and well justified (tied to conceptual underpinnings)?
- *Cognitive Demand*: To what extent do classroom interactions create and maintain an environment of intellectual challenge?
- *Agency*: To what extent do students have opportunities to conjecture, explain, and argue, thus, developing agency and authority?
- *Use of Student Contributions*: To what extent is reasoning elicited, challenged, and refined?
- *Equitable Access*: To what extent do activity structures invite and support active engagement from a diverse range of students?

This framework is highly suitable for our research strategy because it integrates generic and mathematics-related aspects and has a high validity achieved in qualitative projects and professional development. As it emerged within the Diversity in Mathematics Education Center for Learning and Teaching (DIME, 2007), it was also optimized for underprivileged students with its specific focus on *Agency* and *Equitable Access*. Additionally, the declared learning goal is "robust understanding," in other words, the ability "to be effective at dealing with verbally presented, situationally based problems" (Schoenfeld, 2013, p. 609), which resonates well with our focus on developing conceptual understanding in language-responsive ways, much more than frameworks such as the MDI framework, which was optimized for more procedural South African classroom cultures (Adler & Ronda, 2015).

When adapting the framework with respect to the diversity factor in view, language proficiency, we build upon a conceptualization of language proficiency comprising not only lexical and syntactical features but also students' ability to engage in rich discourse practices such as explaining meanings and describing mathematical structure (Moschkovich, 2015). Only slight adaptations of the quality dimensions were needed to incorporate the state of research on language-responsive teaching (Erath et al., 2021), as presented in Fig. 1 (with the core questions from Schoenfeld, 2013, p. 616, and adaptations in grey). Two dimensions were added to capture phenomena often shown as relevant for language-responsive mathematics learning (Prediger & Neugebauer, 2021):

- *Discursive Demand.* Qualitative studies have shown that maintaining *Cognitive Demands* (as captured by the second dimension) often co-occurs with discursive demand (whether students engage in rich discourse practices such as arguing and explaining; Herbel-Eisenmann et al., 2011). However, it seems worth splitting both dimensions in order to capture subtle differences between collective thinking processes and the (individually or collectively conducted) discourse practices—such as reporting a procedure, explaining the meaning of a concept, and arguing—in which explaining and arguing have been shown to be richer and more difficult for students than reporting procedures (Moschkovich, 2015).
- *Connecting Registers.* Developing language and conceptual understanding can be strengthened by multiple multimodal representations (Zahner et al., 2012) and multiple language registers, (i.e. everyday, academic, or formal language). Here, the degree to which representations and registers are not only juxtaposed but deliberately connected is crucial (Adler & Ronda, 2015).

In each dimension (depicted in Fig. 1), teaching practices in 5-min periods are rated on three levels of sophistication: For example, in the dimension of *Mathematical Richness*, discussions are rated as basic if they are "purely rote OR disconnected or unfocused OR consequential mistakes are left unaddressed" (Schoenfeld, 2013, p. 615). They are rated as proficient if "mathematics discussed is relatively clear and current, BUT connections between are either cursory or lacking" (Schoenfeld, 2013, p. 615). A rating of distinguished is given when these connections occur. Raters work from the beginning to the end, so the ratings can also take into account the sequence of 5-min sections (e.g. for *Equitable Access*, we consider the children involved in longer time spans).



	Mathematical Richness	Cognitive Demand	Equitable Access	Agency	Use of Contributions	Discursive Demand	Connecting Registers
Dimensions	To what extent is the mathematics discussed clear, correct, and well justified (tied to conceptual underpinnings)?	To what extent do classroom interactions create and maintain an environment of intellectual challenge?	To what extent do activity structures invite and support active engagement from the diverse range of students?	To what extent do students have opportunities to conjecture, explain, and argue, thus to developing agency and authority?	To what extent is student reasoning elicited, challenged, and refined?	To what extent do students engage in rich discourse practices? (additional dimension)	To what extent are language registers and representations systematically and explicitly connected? (additional dimension)
Level 0	The content is purely rote OR disconnected or unfocused OR consequential mathematical errors or language inaccuracies are not addressed.	Classroom activities are structured so that students mostly apply familiar procedures or memorized facts.	Classroom management is problematic to the point where the lesson is disrupted, OR a significant number of students appear disengaged and there are no overt mechanisms to support engagement.	The teacher initiates conversations. Students' speech turns are short (one sentence or less) and shaped or constrained by what the teacher says or does.	The teacher may note student answers or work, but the student reasoning is not surfaced or pursued. Teacher actions are limited to corrective feedback or encouragement.	No explicit demands to verbalize own ways of thinking, procedures, or solutions OR students only report their processes of calculation.	The content is primarily addressed in only one register/representation OR Different registers/representations are juxtaposed but not related to each other.
Level 1	The content is relatively clear and correct BUT connections between procedures/calculation strategies, concepts, possibly contexts, and the meaning-related language are either *limited or superficial*	Classroom activities offer possibilities of conceptual or language richness or problem-solving challenge, BUT teaching interactions tend to "scaffold away" the challenges and mostly limit students to providing short responses to teacher prompts.	*The participation of students is evenly distributed or the teacher gives support so that a variety of students can participate in activities* *BUT* *the students do not necessarily carry out higher order activities related to content.*	Students have a chance *to talk about mathematical content, their own ideas, and* meaning-related interpretations BUT "the student proposes, the teacher disposes": class discussions and student ideas are not explored or built upon.	The teacher refers to student's thinking and student's meaning-related language, perhaps even to common mistakes BUT *ideas with learning potential are not taken as a basis or problematic ideas are not used as challenges.*	Students are explicitly asked or are used to explaining meanings and justifying concepts, their own ways of thinking, procedures, and solutions BUT formal and meaning-related language resources are not or incorrectly linked with each other.	Content or tasks are translated into another representation/register BUT changes are always conducted in the same direction.
Level 2	The content is relatively clear and correct AND connections between procedures/*strategies*, concepts, *contexts* and meaning-related language are addressed and explained	The teacher's hints or scaffolds *encourage and* support students in "productive struggle" in building understanding and engaging in mathematical practices or language issues. *AND* *Level of demand is maintained by appropriate scaffolds or prompts.*	The teacher actively supports (and to some degree achieves) broad and meaningful participation OR What appear to be established participation structures result in such participation.	Students put forth and defend their ideas and use terminology or meaning-related language. Teacher may ascribe ownership for students' ideas in exposition, *OR* students respond to and build on each other's ideas.	The teacher solicits student thinking and individual use of meaning-related language AND subsequent instruction responds to those ideas by building on productive beginnings or emerging misunderstanding *or language errors.*	Students are explicitly asked or are used to explaining meanings and justifying their own ways of thinking, procedures, and solutions AND formal and meaning-related language resources are correctly related.	The explicit connection between several registers/representations is stimulated AND realized by verbalizing the connection. OR Changes are conducted flexibly in different directions.

Fig. 1 L-TRU framework: language-responsive mathematics teaching for robust understanding (Prediger & Neugebauer [2021]. Adaptations from Schoenfeld's TRU [2013] are marked in grey when they concern language. Adjustments in italics were made for avoiding ceiling effects in our data set)

Methodological Framework of the Quality Video Study

Fig. 2 provides an advanced organizer on the design by which we investigated the impact of quality teaching practices on students in 18 Grade 7 classes ($n = 367$). They were all taught by their regular mathematics teachers and all used the same

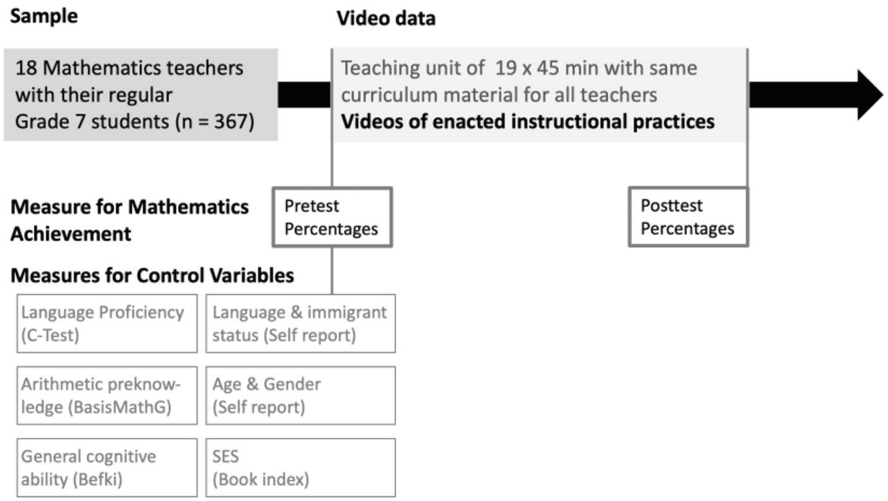

Sample
18 Mathematics teachers with their regular Grade 7 students (n = 367)
Video data
Teaching unit of 19 x 45 min with same curriculum material for all teachers
Videos of enacted instructional practices
Measure for Mathematics Achievement
Pretest Percentages
Posttest Percentages
Measures for Control Variables
Language Proficiency (C-Test)
Language & immigrant status (Self report)
Arithmetic preknow-ledge (BasisMathG)
Age & Gender (Self report)
General cognitive ability (Befki)
SES (Book index)

Fig. 2 Overview of research design for the quality video study

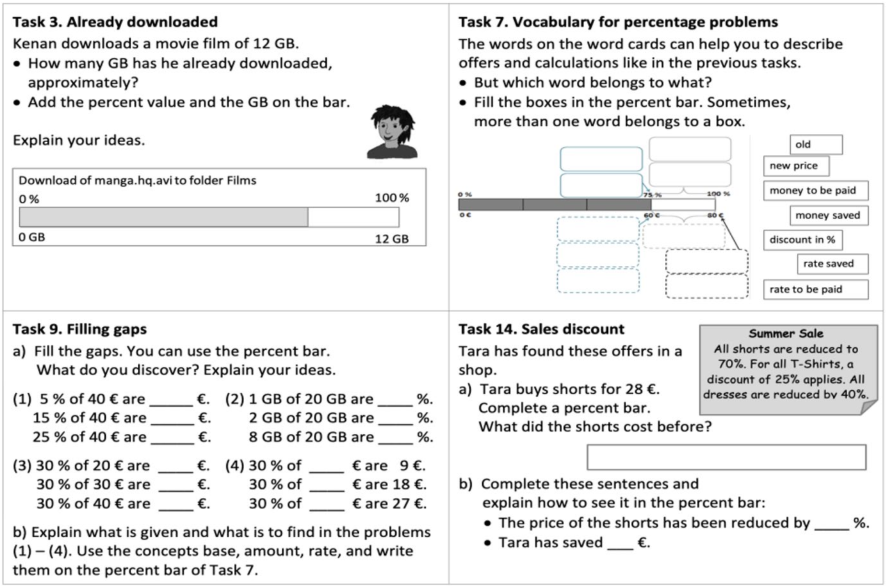
language-responsive curriculum resources aimed at developing a robust conceptual understanding of percentages.

Research Context: Language-Responsive Curriculum Resources for Percentages and Their Proven Effectiveness

All classes involved in this study used curriculum resources for developing students' robust conceptual understanding of percentages in Grade 7 in about 19 lessons of 45 min each (Pöhler & Prediger, 2015).

In the 21 tasks (four are shown in Fig. 3), the curriculum resources (including a teachers' manual) provide a great deal of support for teachers to realize quality instruction in general and in language-responsive classrooms in particular, in four of the seven dimensions:

- *Mathematical Richness* is supported by tasks aiming at developing conceptual understanding along a content trajectory. This consists of several steps, starting from students' informal experiences, eliciting students' strategies for exploiting the relations of involved concepts, and then schematizing into more formal procedures (adapted from van den Heuvel-Panhuizen, 2003). The percent bar serves as the crucial model for understanding the relationships between mathematical concepts base, amount, percentages, and reductions, for instance, when discussing informal strategies determining the base or amount (e.g. in Tasks 3 and 14 in Fig. 3).
- *Cognitive Demands* are provided by opportunities for productive struggle (by different questions in Task 14b, addressing the same percent bar) and by working with patterns or the use of inverse structures (Task 9).



Task 3. Already downloaded

Kenan downloads a movie film of 12 GB.

- How many GB has he already downloaded, approximately?
- Add the percent value and the GB on the bar.

Explain your ideas.

Download of manga.hq.avi to folder Films

0 %
100 %
0 GB
12 GB

Task 7. Vocabulary for percentage problems

The words on the word cards can help you to describe offers and calculations like in the previous tasks.

- But which word belongs to what?
- Fill the boxes in the percent bar. Sometimes, more than one word belongs to a box.

Task 9. Filling gaps

a) Fill the gaps. You can use the percent bar.
What do you discover? Explain your ideas.

(1) 5 % of 40 € are _____ €.
15 % of 40 € are _____ €.
25 % of 40 € are _____ €.

(2) 1 GB of 20 GB are ____ %.
2 GB of 20 GB are ____ %.
8 GB of 20 GB are ____ %.

(3) 30 % of 20 € are ____ €.
30 % of 30 € are ____ €.
30 % of 40 € are ____ €.

(4) 30 % of ____ € are 9 €.
30 % of ____ € are 18 €.
30 % of ____ € are 27 €.

b) Explain what is given and what is to find in the problems (1) – (4). Use the concepts base, amount, rate, and write them on the percent bar of Task 7.

Task 14. Sales discount

Tara has found these offers in a shop.

Summer Sale
All shorts are reduced to 70%. For all T-Shirts, a discount of 25% applies. All dresses are reduced by 40%.

a) Tara buys shorts for 28 €.
Complete a percent bar.
What did the shorts cost before?

b) Complete these sentences and explain how to see it in the percent bar:

- The price of the shorts has been reduced by ____ %.
- Tara has saved ___ €.

Fig. 3 Tasks from the language-responsive curriculum resources on percentages

- *Connecting Registers* is initiated by the permanent use of the visual model percent bar (in almost all tasks) and by connecting students' everyday language to the meaning-related language (e.g. part of the whole, new price, and old price) and the technical language (e.g. rate, base and amount in Tasks 7 and 14).
- *Discursive Demands* are intended when students are invited to report their informal strategies (Task 3) or describe regularities (Task 9). Due to the empirically proven relevance of rich discourse practices as epistemic catalysts (Moschkovich, 2015), learners benefit from explaining meaning with meaning-related vocabulary (Task 14), and for this, elicit, systematize (Task 7), and practice (Task 9) meaning-related vocabulary. Beyond the tasks, we presume that discursive demands unfold in teaching practices.

Although these four quality dimensions can be supported by curriculum resources, their enactment in teaching practices is always crucial for maintaining the demands (Stein et al., 2000). The classes working with these curriculum resources were videotaped within a larger project that conducted a controlled field trial that initially had 655 students. Overall, the curriculum resources showed effectiveness when comparing the learning gains of the intervention classes (working with the curriculum resources) to the control-group classes (working with their regular textbooks with less support for establishing quality instruction): The ANOVA revealed significant differences (with $F_{\text{time} \times \text{group}}$ (1, 653) = 20.74, $p < 0.001$, $\eta^2 = 0.011$), with higher differences for robust understanding than for basic understanding (Prediger & Neugebauer, 2022).

However, remarkably large differences occurred between the intervention classes: with an intraclass correlation coefficient (ICC) of 0.30, 30% of the variance can be explained by class adherence. This large ICC occurred although all teachers in the intervention classes were prepared in professional development (four sessions of 3 h each), received supportive curriculum resources, and realized them with a workbook completion rate of more than 75%. As composition effects did not occur, and school effects were not controlled, we decided to conduct the current quality video study because it became necessary to investigate how the quality of enacted teaching practices influenced mathematics achievement.

Measures for Mathematics Achievement and Control Variables

- For measuring the targeted mathematics achievement in robust understanding of percentages, a standardized posttest was conducted. This percent test assesses conceptual understanding and flexible use of percentages (Pöhler et al., 2017). It took 40 min and consisted of open items in three problem types: "find the amount," "find the base," and "find the base after reduction." For each of them, items varied in three formats: "pure format," "text format," and "visual format," with percent bar representations. The test has a satisfactory internal consistency for its 29 items, with Cronbach's $\alpha = 0.834$ for the posttest (in a sample of $n = 1120$ students). For the control variables, the following measures were administered before the teaching unit and served as relevant control variables in the multilevel models:
- German academic language proficiency was assessed by a *C*-test, a widely used, economical, and valid measure based on cloze texts (Grotjahn et al., 2002). It took 15 min and consisted of three texts in everyday and academic formal language. It reached a consistency of Cronbach's $\alpha = 0.774$ ($n = 1122$).
- Arithmetic preknowledge relevant for learning percentages (fractions, mental models for multiplication and division, proportional reasoning, etc.) was assessed by a standardized test with Cronbach's $\alpha = 0.83$ (35 min, 28 items, $n = 1120$).
- Percentage preknowledge was captured by visual and text formats of "find the amount" (maximum score of 6). Only three items were selected from the posttest (given the little expectable knowledge on percentages before the teaching unit), so it took only 6 min, but internal consistency is limited.

Further, control variables were captured to control for the comparability of the video sample with the full sample but not for the multilevel analysis:

- Age, gender, multilingual family socialization (operationalized by languages spoken with parents or siblings), and immigration status (operationalized by parents' and own country of birth) were captured in a students' self-report questionnaire.
- A subconstruct of general cognitive ability, fluid intelligence, was measured using a matrix test for fluid intelligence. The internal consistency was Cronbach's $\alpha = 0.763$ ($n = 1124$).

Sample

Among the teachers of intervention classes who worked with the language-responsive curriculum resources (see subsection on research context), 26 teachers volunteered to be videotaped with two cameras: one showing the entire classroom and one following the teacher. For 18 of the videotaped classes from various schools, all measures on mathematics achievement and control variables were completed; the video data from these classes form the data corpus of the current video study. The 18 teachers in these classes held teaching certificates in mathematics (middle school or high school) and had 1 to 18 years of teaching experience. All were introduced to the ideas of conceptually oriented and language-responsive teaching in three professional development sessions of 3–5 h each. The 18 classes had a median of 22 students, with 367 students completing all tests. For this quality video study sample, a positive selection bias might be assumed due to teachers' voluntary participation. However, Table 1 reveals that the quality video study sample started with comparable abilities as the initial full sample of all intervention classes (Prediger & Neugebauer, 2022): No significant difference occurred in language proficiency, arithmetic preknowledge, percentage preknowledge, or mathematics achievement in the percent posttest.

Table 1 Descriptive data for the initial full sample and the sample of the quality video study

Variables with m (SD) or distribution (in %)	Initial full sample (intervention group in controlled trial $n = 587$)	Sample of the quality video study ($n = 367$)	t-tests/χ^2 tests for significance of differences
Mathematics achievement (posttest score, max. 29)	12.81 (06.38)	12.32 (5.86)	$p = 0.30$
Key control variable			
Language proficiency (max. 60)	38.77 (11.88)	37.00 (12.2)	$p = 0.10$
Arithmetic preknowledge (max. 63)	34.52 (13.87)	35.16 (13.52)	$p = 0.53$
Percentage preknowledge (max. 6)	2.52 (01.86)	1.6 (01.89)	$p = 0.31$
Further control variables			
General cognitive ability (max. 16)	9.65 (03.29)	9.91 (03.40)	$p = 0.21$
Age (in years)	12.78 (00.77)	12.71 (0.75)	$p = 0.64$
Multilingual background (multi-/monolingual)	49%/51%	59%/41%	$p = 0.46$
Gender (female/male)	51%/49%	54%/46%	$p = 0.78$
Immigration status (born in Germany/immigrated)	92%/8%	89%/12%	$p = 0.65$

Methods for Rating Enacted Teaching Practices in the L-TRU Framework

The methods of rating the quality of the videotaped enacted teaching practices in the L-TRU framework (see Fig. 1) followed five steps (detailed in Prediger & Neugebauer, 2021):

Step 1. Segmenting and Initial Rating with the Original Scale. Each videotaped lesson was split into segments of up to 5 min, starting new segments with every change of activity type (whole-class discussion, student presentation, small group work, and individual work). and every new task. Each of the 497 segments was rated on a 3-point scale (score 0 = basic, 1 = proficient, and 2 = distinguished). To ensure reliability, a rating protocol with flowcharts was iteratively refined in the initial coders' collective discussions.

Step 2. Adjusting the Scales. Figure 4 shows how many 5-min segments were scored 0, 1, and 2 for the original scales (which followed the operationalization of Schoenfeld, 2013, for the first five dimensions). This reveals substantial ceiling effects in the data set, as 80% of the rated 5-min segments in the video quality study data corpus were rated by a 2 with respect to *Mathematical Richness* and *Cognitive Demand*. Indeed, these high floor percentages reveal that the curriculum resources succeeded in supporting teachers' enactment of high-quality teaching in the majority of the time and in the majority of dimensions. However, to investigate the impact of high-quality teaching on students' learning with fewer ceiling effects, the scales were adjusted by the conditions already included in italics in Fig. 2.

Step 3. Determining the Interrater Reliability of the Adjusted Scales. For the adjusted scales, the interrater reliabilities were determined with two independent ratings on 1610 min (2/3 of the total) of video material. The ratings never deviated more than one point. The left part of Table 2 shows that Cohen's kappa

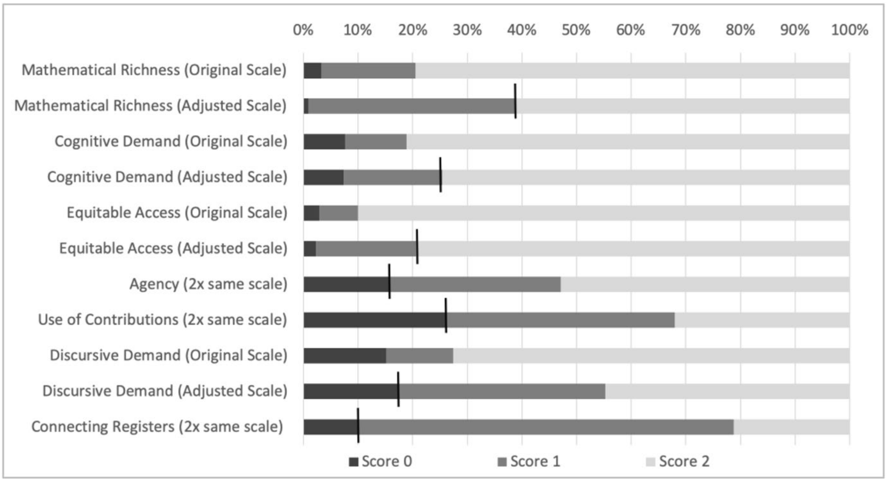


Fig. 4 Distribution of rated 5-min segments in the original scales with ceiling effects (step 2) and in adjusted scales (step 3) and black bars marking the cutoffs for dichotomization (step 4)

Table 2 Interrater reliability and correlation between the adjusted quality dimensions

	Interrater reliability	Correlation between 5-min segments scored as high/low with respect to cutoffs						
	Cohen’s kappa κ	*Mathematical Richness*	*Cognitive Demand*	*Equitable Access*	*Agency*	*Use of Contributions*	*Discursive Demand*	*Connecting Registers*
Mathematical Richness	0.86	1	0.38	0.26	0.17	0.30	0.30	0.12
Cognitive Demand	0.81		1	0.35	0.34	0.27	0.33	0.21
Equitable Access	0.88			1	(0.34)	(0.25)	0.25	(0.20)
Agency	0.73				1	0.41	0.39	0.30
Use of Contributions	0.69					1	(0.41)	(0.16)
Discursive Demand	0.75						1	0.50
Connecting Registers	0.79							1
Overall	0.78	(nonsignificant correlations with $p > 0.5$ in brackets)						

ranges from 0.69 (sufficient) to 0.86 (good) in all dimensions, with an overall κ = 0.78. After that, all moments of disagreement were solved by consensus so the successive analyses could be based on 100% rater agreement.

Step 4. Dichotomizing the Scales. To convert multiple pieces of information about 5-minute segments into one judgment per dimension, scales were dichotomized and metric variables were derived as follows. For each dimension, a cut-off was chosen so that segments with high/low scores comprised more than 14% of the coded segments each. Figure 4 shows the cut-offs using black bars: For *Mathematical Richness*, *Cognitive Demand*, and *Equitable Access*, the cut-off was set between 1 and 2 points, while for *Agency*, *Use of Contributions*, *Discursive Demand*, and *Connecting Registers*, the cut-off was set between 0 and 1 point. Table 2 also documents that the correlation of segments rated as high or low is between 0.12 and 0.50, which confirms the intended characteristic that the seven dimensions are not independent of each other, but nevertheless cover different aspects (dependencies and differences are discussed, also using case studies, in Prediger & Neugebauer, 2021).

Step 5. Deriving Metric Variables for Quality Degrees. The distribution of rated segments in each class was transformed into a metric variable "degree of …" by counting the percentages of segments on a high score. For example, the degree of *Mathematical Richness* in a class determines the rate of 5-min segments above the cutoff among all 5-min segments of that class. Table 3 presents the mean and the standard deviation of these derived metric degree variables. Their standard deviation between 0.20 and 0.28 shows a comparable spread for all degree variables, so the prerequisites of the multilevel analysis are satisfied.

Table 3 Metric variable descriptives: definition, mean, and standard deviation for the seven derived metric variables of quality degrees in each class

	Definition	Descriptive statistics	
Quality degree of …	Operationalization of degrees from high/low scores	*m*	(SD)
Mathematical Richness	What percent of segments is rated 2 points?	0.46	(0.26)
Cognitive Demand	What percent of segments is rated 2 points?	0.67	(0.28)
Equitable Access	What percent of segments is rated 2 points?	0.82	(0.21)
Agency	What percent of segments is rated 1 or 2 points?	0.44	(0.21)
Use of Contributions	What percent of segments is rated 1 or 2 points?	0.32	(0.26)
Discursive Demand	What percent of segments is rated 1 or 2 points?	0.26	(0.26)
Connecting Registers	What percent of segments is rated 1 or 2 points?	0.21	(0.20)

Methods for the Multilevel Analysis

To determine the connection between the quality degrees of teaching practices and mathematics achievement, simple correlations would not be adequate as they do not account for class effects and individual differences in students' abilities; in particular, the sample with a high ICC of 0.30 demands the consideration of the multilevel structure. We therefore used a multilevel analysis, "which is a methodology for the analysis of data with complex patterns of variability, with a focus on nested sources of such variability" (Snijders & Bosker, 2012, p. 1). In our case, the nesting stems from students being nesting in classes with equal quality degrees.

Multilevel analyses (also called hierarchical linear models, mixed models, or random coefficient models) address the variability between measured variables on each level because only addressing one level could reveal incorrect inferences such as ecological biases (erroneously referring collective data to individuals) or atomistic biases (erroneously referring individual data to the collective; Hox et al., 2018). Student variables of language proficiency, arithmetic preknowledge, and percentage preknowledge are treated as level 1 data in our model. The seven quality degrees of the videotaped classrooms are treated as level 2 data, whereas teacher and school data were not available. The analytic focus of the multilevel analysis is the predictive power of level 2 variables (quality degrees) for mathematics achievement (measured by tests on percentages) when controlling for the level 1 student variables (Enders & Tofighi, 2007). The student variables on level 1 are centered at the grand mean, which allows interpretation of the intercept (Snijders & Bosker, 2012, p. 71) and avoids issues of convergence (Hox et al., 2018, pp. 49–50). The selected student variables on level 1 are standardized so that the estimated *b*-weights in the models can easily be compared and interpreted as predicting changes in the raw score of the posttest. Since full maximum likelihood models require 30 classes with 30 individuals to avoid biased parameter estimates, we chose restricted maximum likelihood models, which Hox and McNeish (2020) showed in simulations can still provide reliable findings on random shares of variance for sample sizes with 7–10 classes. Both models revealed comparable results but the chosen restricted maximum likelihood models with more reliability for our 18 classes.

Results on Achieved Quality Degrees and Their Interaction with Achievement

In the research question, we asked which kinds of interactions between the quality of teaching practices and students' abilities (in language proficiency) create learning opportunities (for a robust understanding of percentages) when the curriculum resources are held constant (with a focus on enhancing the mathematics learning of students with diverse language proficiency). Answering this question required first investigating which quality degrees of teaching practices can be reached using a carefully designed language-responsive curriculum resource (summarized in the first subsection) and then a multilevel analysis (presented in the second subsection).

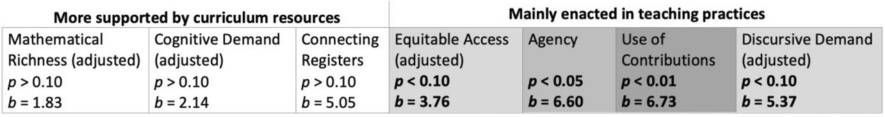
Achieved Quality in the Seven Dimensions

Prior to the study, we assumed that three dimensions, *Mathematical Richness*, *Cognitive Demand*, and *Connecting Registers*, can be more supported by the curriculum resources, whereas the other four, *Equitable Access*, *Agency*, *Use of Contributions*, and *Discursive Demand*, might depend more on the enactment in the teaching practices. Indeed, Fig. 5 (which has already been presented in the Methods section, as further methodological decisions about the cutoffs had to be justified using it) shows that the support provided by the curriculum resources (and the professional development course) resulted in remarkably high scores: The videorecorded teachers spent 79% of their time working on distinguished levels of *Mathematical Richness* (2 points). This means that the content was relatively clear and correct and that connections between strategies and meaning-related language were addressed and explained (see Fig. 1). Also, in 81% of the 5-min segments, the teaching practices reached distinguished levels of *Cognitive Demand*, meaning that the teachers' hints or scaffolds supported students in productive struggle, building understanding, and engaging in mathematical practices. In order to differentiate better between the teaching practices of the different classes, the two scales were adjusted to even higher quality requirements. For the dimension *Connecting Registers*, teachers spent 69% of their time working on the segments at the 1-point level and 22% at the 2-point level; in other words, 69% of the 5-min segments were spent connecting two representations in a particular way, while those at the 2-point level involving connecting them in multiple ways cannot be expected to have taken up more time.

Although not really supported by the curriculum material, 74% of the segments were rated at 2 for discursive demands and 91% for *Equitable Access*, so both scales were adjusted (now 26% and 81%, respectively, at the adjusted 2-point level). No adjustments were necessary for *Agency* (with 53% of the segments at 2 points) and *Use of Contributions* (with 32% of the segments at 2). In these four dimensions, maximally 17% of the 5-min segments were rated at the 0, so low quality was rare for the observed teachers. However, the reported means and standard deviations (in Table 3) reveal a high variance (between 0.20 and 0.28 for each quality degree) between classes, so the interaction of varying quality degrees in mathematics achievement must be studied.

Interaction of Quality Degrees and Mathematics Achievement

Based on these preliminary inquiries, the research question can be treated by analyzing how the quality degrees of teaching practices interact with students' posttest



More supported by curriculum resources			Mainly enacted in teaching practices			
Mathematical Richness (adjusted)	Cognitive Demand (adjusted)	Connecting Registers	Equitable Access (adjusted)	Agency	Use of Contributions	Discursive Demand (adjusted)
$p > 0.10$	$p > 0.10$	$p > 0.10$	**$p < 0.10$**	**$p < 0.05$**	**$p < 0.01$**	**$p < 0.10$**
$b = 1.83$	$b = 2.14$	$b = 5.05$	**$b = 3.76$**	**$b = 6.60$**	**$b = 6.73$**	**$b = 5.37$**

Fig. 5 Comparison of significant predictors (p values and estimated b-weights from Table 4)

scores when controlling for students' most relevant abilities. Table 4 shows the calculated models for predicting mathematics achievement in terms of the students' outcome of robust understanding. The empty model 0 ($R^2 = 0.304$ for level 2) resonates with the already reported ICC for class adherence. Additionally, the full model 0a was calculated with all quality dimensions at once ($R^2 = 0.538$ for level 1 and $R^2 = 0.643$ for level 2), but due to the interaction between dimensions, it cannot predict the achievement better than multilevel models 1–7 with only one level 2 variable, so it was not included in Table 4. The multilevel models 1–7 each include one of the quality degrees (the metric variable derived for the quality dimension).

On level 1, all seven models show the highly significant predictive power of students' abilities in language proficiency, arithmetic preknowledge, and percentage preknowledge. Due to the grand mean centering, the intercept of 12.2 in the models can be interpreted as the estimated score for the average student (notwithstanding the quality degrees on the class level). An estimated *b*-weight of 3.48 for arithmetic preknowledge means that if a students' arithmetic preknowledge is one standard deviation higher than the mean, the model predicts an additional score of 3.48 in the posttest. Thus, all three of the student abilities that had been hypothesized to interact with the mathematics learning are shown to have this influence, but the impact of the mathematical preknowledge is much higher than that of language proficiency.

On level 2, all seven quality degrees reveal some positive influence of the quality degrees on mathematics achievement. However, only in four models do the quality degrees have significant predictive power for achievement. The additional predictive power is significant on a 10% level for the degrees of *Equitable Access* and *Discursive Demand* (marked with $^{\diamond}$ in Table 4) and significant on a 5% level for *Agency*. The most significant predictor is the *Use of Contributions* (significant on a 1% level). The estimated *b*-weight of 6.73 means that a theoretical increase of the quality degree of *Use of Contributions* by 100% would predict that the estimated score in the posttest increases by 6.73. In other words, an increase of 10% in segments rated at 1 or 2 points rather than 0 predicts an increase of 0.673 in the posttest score, which is 12% of the posttest standard deviation of 5.68.

The high values of R^2 show the explanative power of the models gained even if we only had 18 classes. Whereas model 0 documents R^2 of 0.304 (i.e. 30.4% of the variance in posttest scores is traced back simply to class adherence), Models 1–7 reveal R^2 between 0.627 and 0.646, which means that 62.7% to 64.6% of the variance can be explained by students' diverse abilities combined with the quality dimensions of the implemented teaching practices: These combinations explain twice as much variance as the empty model 0.

Conclusion, Limitations, and Implications

Summary and Discussion of Results

As students with limited access to academic language have often been shown to experience opportunity gaps (Callahan, 2005; Herbel-Eisenmann et al., 2011;

Table 4 Multilevel models for predicting mathematics achievement: comparing seven quality dimensions

	Model 0 *b* (SE)	Model 1 *b* (SE)	Model 2 *b* (SE)	Model 3 *b* (SE)	Model 4 *b* (SE)	Model 5 *b* (SE)	Model 6 *b* (SE)	Model 7 *b* (SE)
Level 1: student variables								
(Intercept)	11.93*** (0.81)	12.21*** (0.51)	12.20*** (0.51)	12.24*** (0.47)	12.24*** (0.46)	12.28*** (0.44)	12.26*** (0.49)	12.17*** (0.47)
Language proficiency		0.65** (0.24)	0.65** (0.24)	0.63** (0.24)	0.64** (0.24)	0.63** (0.24)	0.63** (0.24)	0.66** (0.24)
Arithmetic preknowledge		3.48*** (0.28)	3.51*** (0.29)	3.48*** (0.28)	3.53*** (0.28)	3.48*** (0.28)	3.48*** (0.28)	3.54*** (0.29)
Percentage preknowledge		0.77** (0.24)	0.78*** (0.24)	0.78*** (0.24)	0.78*** (0.24)	0.78*** (0.24)	0.78*** (0.24)	0.77** (0.24)
Level 2: quality degree of								
Mathematical Richness		1.83 (1.92)						
Cognitive Demand			2.14 (1.81)					
Equitable Access				3.76° (1.94)				
Agency					6.60* (2.85)			
Use of Contributions						6.73** (2.50)		
Connecting Registers							5.05 (3.11)	
Discursive Demand								5.37° (2.77)
R^2 (level 1)	-	0.529	0.521	0.547	0.537	0.561	0.544	0.529
R^2 (level 2)	0.304	0.642	0.633	0.642	0.628	0.640	0.646	0.627

(Significance is indicated on several levels: $^{\circ}p < 0.1$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$)

Secada, 1992), this study aimed at investigating how to increase their learning opportunities and enhance their language and mathematics learning for developing conceptual understanding, using the topic of percentages as an example. Following the research agenda promoted by Cai et al. (2020), we investigated the interactions between the quality of teaching practices, mathematics achievement, and students' abilities (in language proficiency, arithmetic preknowledge, and percentage preknowledge) when the curriculum resources are held constant. For this, we used data from a field study in which 18 classes used the same curriculum resources optimized for enhancing mathematics learning with diverse language proficiency (Pöhler & Prediger, 2015), following established language-responsive design principles (Erath et al., 2021).

The preliminary inquiry of achieved quality degrees revealed that indeed, the curriculum resources seem to have supported the teachers to enact high-quality teaching practices that meet the learning needs of students with limited academic language proficiency. Of course, the effect cannot be ascribed solely to the curriculum resources, as we worked with volunteer teachers who also attended an intense PD for enacting the teaching unit in a productive way and have no videos from a control group who did not use the curriculum material. Thus, these findings must not be misinterpreted as proof of simple implementation pathways, as enacted teaching practices can always strengthen or inhibit the effectiveness of curriculum resources (Cohen et al., 2003). However, it is encouraging to have a modest initial indication that the curriculum resources and the PD together seem to have established a high floor of quality teaching practices even in classes with diverse language proficiencies.

With respect to the relevance of students' abilities, the results in Table 4 on Level 1 show the highly significant predictive power of students' language proficiency, arithmetic preknowledge, and percentage preknowledge. Although these findings could be expected based on the existing state of research (Haag et al., 2013; Secada, 1992), the explained variance is remarkably strong with $R^2 = 0.526$, slightly varying in terms of hundredths in models 1–7 (much stronger than, for example, in the primary education study by Decristan et al. (2015), who reported $R^2 = 0.135$ for cognitive ability, subject matter preknowledge, and language proficiency). Even when controlling for mathematical preknowledge, language proficiency still has a (small but significant) influence on achievement, persisting even when controlling for diverse mathematical preknowledge.

Based on these preparations, the main research question of this paper was refined into the analytic task of determining how the quality degrees of teaching practices interact with mathematics achievement in posttest scores (level 2 in Table 4) when controlling for students' most relevant abilities (level 1 in Table 4). Although the seven models all reveal a slightly positive impact of each quality degree on mathematics achievement, this impact is significant only for four quality degrees. The summary of p values and estimated b-weights in Fig. 5 shows that the interaction is not significant on a 10% level for those particular quality degrees that are more supported by the curriculum resources (*Mathematical Richness*, *Cognitive Demand*, and *Connecting Registers*).

The quality dimensions *Mathematical Richness*, *Cognitive Demand*, and *Connecting Registers* have very high ratings with potential ceiling effects. Their spread between the intervention classes does not additionally predict mathematics achievement in a significant way. It is most probable to assume that this null finding relates to the observation that the differences between classes occur on a very high floor. So, the null finding of missing significance for additional predictive power should not be misinterpreted as contradicting other findings about the high relevance of *Mathematical Richness* (Adler & Ronda, 2015; Hill et al., 2008), *Cognitive Demands* (Hill et al., 2008; Praetorius et al., 2018; Stein et al., 2000), and *Connecting Registers* (Adler & Ronda, 2015). When compared to control classes in the controlled trial, they indeed provide evidence for effectiveness (Prediger & Neugebauer, 2022), but the current research design is optimized for searching for even additional effects within the intervention classes.

In contrast, the differences in those quality dimensions that are mainly enacted in the teacher-student interaction of the teaching practices have an additional influence that is nearly significant:

- The quality degree of *Use of Contributions* is a highly significant predictor of student achievement. We interpret the degree of *Use of Contributions* as the indicator for teachers' adaptivity in dealing with students' ideas, and when the *Mathematical Richness*, *Cognitive Demand*, and degree of *Connecting Registers* are already ensured by the curriculum resources, then teachers' *Use of Contribution* is most influential. In this way, dealing with the students' contributions is indirectly scaffolded by the curriculum resources (Cohen et al., 2003), but even on this high floor, it still depends massively on how teachers notice students' ideas (Empson & Jacobs, 2008), and the additional differences become highly significantly predictive for student achievement. These results confirm general findings (Seidel & Shavelson, 2007) that interactional components are more powerful predictors for learning gains than structural elements.
- *Equitable Access*, which shows the highest mean (m = 0.82 even after adjustment of the scale), still has a significant influence on the 10% level, which is remarkable, as the high floor might have resulted in no additional predictive power. The result confirms the findings by Ing et al. (2015) that posttest achievements are highly influenced by the level of student participation. Again, this is hardly scaffolded by the curriculum resources.
- The same applies to *Agency*. The impact of this dimension resonates with many qualitative studies emphasizing *Agency* as an entry point for mathematical reasoning (DIME, 2007; Herbel-Eisenmann et al., 2011).
- *Discursive Demand* also has a significant influence on the 10% level. These findings support Herbel-Eisenmann et al.'s (2011) case studies showing how students' productive struggle presupposes being engaged in rich discourse practices.

Conclusion and Implications

In total, we can summarize our theoretical and empirical contribution to the general question "What kinds of interactions among tasks, teaching, and students create learning opportunities for a specific learning goal?" (Cai et al., 2020, p. 16): For the specific learning goal of conceptual understanding of percentages, we aimed at disentangling the role of curriculum resources and teaching practices. Eighteen classes were filmed, all taught with the same language-responsive curriculum resources that meet the needs of students with low language proficiency. The rating of the teaching practices in seven quality dimensions reveals that all classes operated on such a very high level of *Mathematical Richness*, *Cognitive Demands*, *Equitable Access*, and *Discursive Demand* that the scales even had to be adjusted (Fig. 5). A humble indicator of their effectiveness had already been found, as control classes working without the language-responsive curriculum resources had significantly lower student achievement in the posttest (Prediger & Neugebauer, 2022).

However, among these high-quality intervention classrooms, the quality degrees of the teaching practices still varied, and this variation can additionally predict students' achievement for those quality dimensions that are mainly enacted in teaching practices of *Equitable Access*, *Agency*, *Use of Contributions*, and *Discursive Demand*. Once the curriculum resources are designed in a way that meets language needs, a high-quality floor for teaching practices can be established, and all students profit from the learning opportunities in a significant way. Even if these findings must be treated with some caution due to some methodological limitations (discussed in the next subsection), they provide interesting first indications that those quality dimensions that are hard to support using curriculum materials have an additional impact on students' mathematics learning. Beyond the often-shown potentials for improving language-responsive mathematics teaching and learning by the design of curriculum resources (de Araujo et al., 2018; Zahner et al., 2012), the current results indicate that classroom enactment might play an additional role, as was expected from multiple qualitative findings (Herbel-Eisenmann et al., 2011). This calls for future studies with more detailed inquiries into disentangling the impact of curriculum resources and teaching practices quantitatively, and also in experimental trials with and without curriculum resources.

In the current state of research, the presented findings should also have implications for maximizing all students' learning opportunities by the design of curriculum resources and professional development projects: Good curriculum resources matter! Therefore, the international mathematics education research community together with local authorities should develop curriculum resources that support teachers to establish and maintain *Mathematical Richness*, *Cognitive Demands*, and *Connecting Registers*, in particular for students with low language proficiency, although all students can profit from this. On this high floor, teachers' interaction with students can additionally maximize the learning opportunities, in particular with respect to the *Use of Students' Contributions*, which requires professional development sensitizing teachers for noticing students' ideas (Empson & Jacobs, 2008) and students' language (Moschkovich, 2015).

Limitations and Future Research

Of course, this study is only a small contribution to the big question of how to maximize learning opportunities, and its findings must be interpreted with respect to its methodological limitations:

First, although the sample size of 18 classrooms is methodologically acceptable for multilevel analysis with restricted maximum likelihood (Hox & McNeish, 2020), future studies should extend the number of involved classes in order to further strengthen the stability of findings.

Second, all observed classes were taught by volunteer teachers with intense professional development that supported them in enacting the potentials of the curriculum resources in a productive way, so the effects were not produced by the curriculum resources alone (Cohen et al., 2003). In future research, we intend to also compare control classes working without the given curriculum resources and with teachers who only received the curriculum resources but not professional development. Findings about the relevance of professional development and curriculum resources can then be disentangled more systematically.

Acknowledgements This study was conducted as part of the project MuM-innovation, supported by the National Ministry of Education and Research (under BMBF-Grant 03VP02270 to S. Prediger) in the research framework of DZLM, the German Center for Mathematics Teacher Education (funded 2011–2020 by German Telekom Foundation). The multilevel analysis was supported by a cooperation project with the IPN|Leibniz Institute for Science and Mathematics Education, to which the second author belongs. The authors thank Alan Schoenfeld for the inspiring conversation about the L-TRU framework and the interpretation of this paper's results.

Funding Open Access funding enabled and organized by Projekt DEAL.

References

Adler, J., & Ronda, E. (2015). A framework for describing mathematics discourse in instruction and interpreting differences in teaching. *African Journal of Research in Mathematics, Science and Technology Education, 19*(3), 237–254. https://doi.org/10.1080/10288457.2015.1089677

Barwell, R., Clarkson, P., Halai, A., Kazima, M., Moschkovich, J., Planas, N., & Villavicencio, M. (Eds.). (2016). *Mathematics education and language diversity: The 21st ICMI Study*. Springer.

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

Cai, J., Morris, A., Hohensee, C., Hwang, S., Robison, V., Cirillo, M., Kramer, S., 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

Callahan, R. M. (2005). Tracking and high school English learners: Limiting opportunity to learn. *American Educational Research Journal Summer, 42*(2), 305–328. https://doi.org/10.3102/00028312042002305

Charalambous, C. Y., & Praetorius, A.-K. (2018). Studying mathematics instruction through different lenses: Setting the ground for understanding instructional quality more comprehensively. *ZDM – Mathematics Education, 50*(3), 355–366. https://doi.org/10.1007/s11858-018-0914-8

Cohen, D. K., Raudenbush, S. W., & Ball, D. L. (2003). Resources, instruction, and research. *Educational Evaluation and Policy Analysis, 25*(2), 119–142. https://doi.org/10.3102/01623737025002119 de Araujo, Z., Roberts, S. A., Willey, C., & Zahner, W. (2018). English learners in K–12 mathematics education: A review of the literature. *Review of Educational Research, 88*(6), 879–919. https://doi.org/10.3102/0034654318798093

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

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

Empson, S. B., & Jacobs, V. J. (2008). Learning to listen to children's mathematics. In T. Wood & P. Sullivan (Eds.), *International Handbook of Mathematics Teacher Education* (Vol. 1, pp. 257–281). Sense.

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multi-level models: A new look at an old issue. *Multi-level Modelling Newsletter, 16*(2), 3–9. https://doi.org/10.1037/1082-989X.12.2.121

Erath, K., Ingram, J., Moschkovich, J., & Prediger, S. (2021). Designing and enacting instruction that enhances language for mathematics learning – A review of the state of development and research. *ZDM – Mathematics Education, 53*(2), 245–262. https://doi.org/10.1007/s11858-020-01213-2

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.

Haag, N., Heppt, B., Stanat, P., Kuhl, P., & Pant, H. A. (2013). Second language learners' performance in mathematics: Disentangling the effects of academic language features. *Learning and Instruction, 28,* 24–34. https://doi.org/10.1016/j.learninstruc.2013.04.001.

Herbel-Eisenmann, B., Choppin, J., Wagner, D., & Pimm, D. (2011). *Equity in Discourse for Mathematics Education*. Springer.

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

Hox, J. J., & McNeish, D. (2020). Small samples in multilevel modeling. In R. Van de Schoot & M. Miočević (Eds.), *Small sample size solutions* (pp. 215–225). Routledge.

Hox, J. J., Moerbeek, M., & van de Schoot, R. (2018). *Multilevel Analysis*. Routledge.

Ing, M., Webb, N. M., Franke, M. L., Turrou, A. C., Wong, J., Shin, N., & Fernandez, C. H. (2015). Student participation in elementary mathematics classrooms: The missing link between teacher practices and student achievement? *Educational Studies in Mathematics, 90*(3), 341–356. https://doi.org/10.1007/s10649-015-9625-z

Kilpatrick, J. (2003). What works? In S. L. Senk & D. R. Thompson (Eds.), *Standards' based school mathematics curricula* (pp. 471–493). Lawrence Erlbaum.

Kilpatrick, J., Swafford, J., & Findel, B. (Eds.). (2001). *Adding it up: Helping children learn mathematic*s. National Academy Press.

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

Moschkovich, J. (2010). Language(s) and learning mathematics: Resources, challenges, and issues for research. In J. Moschkovich (Ed.), *Language and mathematics education: Multiple perspectives and directions for Research* (pp. 1–28). Information Age. https://doi.org/10.1016/j.jmathb.2015.01.005

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

Parker, M., & Leinhardt, G. (1995). Percent: A privileged proportion. *Review of Educational Research, 65*(4), 421–481. https://doi.org/10.3102/00346543065004421

Pöhler, B., George, A.-C., Prediger, S., & Weinert, H. (2017). Are word problems really more difficult for students with low language proficiency? Investigating percent items in different formats and types. *International Electronic Journal of Mathematics Education*, *12*(3), 667–687. https://doi.org/10.29333/iejme/641

Pöhler, B., & Prediger, S. (2015). Intertwining lexical and conceptual learning trajectories: A design research study on dual macro-scaffolding towards percentages. *Eurasia Journal of Mathematics, Science and Technology Education*, *11*(6), 1697–1722. https://doi.org/10.12973/eurasia.2015.1497a

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

Prediger, S., & Neugebauer, P. (2021). Capturing teaching practices in language-responsive mathematics classrooms: Extending the TRU framework "teaching for robust understanding" to L-TRU. *ZDM – Mathematics Education*, *53*(2), 289–304. https://doi.org/10.1007/s11858-020-01187-1

Prediger, S., & Neugebauer, P. (2022, online first). Can students with different language backgrounds equally profit from a language-responsive instructional approach for percentages? Differential effectiveness in a field trial. *Mathematical Thinking and Learning*, 1–21. https://doi.org/10.1080/10986065.2021.1919817

Prediger, S., Wilhelm, N., Büchter, A., Gürsoy, E., & Benholz, C. (2018). Language proficiency and mathematics achievement – Empirical study of language-induced obstacles in a high stakes test, the central exam ZP10. *Journal für Mathematik-Didaktik*, *39*(Supp. 1), 1–26. https://doi.org/10.1007/s13138-018-0126-3

Schoenfeld, A. H. (2013). Classroom observations in theory and practice. *ZDM – Mathematics Education, 45*(4), 607–621. https://doi.org/10.1007/s11858-012-0483-1

Schoenfeld, A. H., Floden, R., Chidiac, F. E., Gillingham, D., Fink, H., Hu, S., Sayavedra, A., Weltman, A., & Zarkh, A. (2018). On classroom observations. *Journal for STEM Education Research, 1*(1–2), 34–59. https://doi.org/10.1007/s41979-018-0001-7

Secada, W. G. (1992). Race, ethnicity, social class, language and achievement in mathematics. In D. A. Grouws (Ed.), *Handbook of Research on Mathematics Teaching and Learning* (pp. 623–660). MacMillan.

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

Snijders, T. A. B., & Bosker, R. (2012). *Multi-level analysis: An introduction to basic and advanced multi-level modeling* (2nd ed.). Sage.

Stein, M., Smith, M. S., Henningsen, M. A., & Silver, E. A. (2000). *Implementing standards-based mathematics instruction*. In *A casebook for professional development*. Teachers College Press.

Van den Heuvel-Panhuizen, M. (2003). The didactical use of models in Realistic Mathematics Education: An example from a longitudinal trajectory on percentage. *Educational Studies in Mathematics, 54*(1), 9–35. https://doi.org/10.1023/B:EDUC.0000005212.03219.dc

Wilhelm, A. G., Munter, C., & Jackson, K. (2017). Examining relations between teachers' explanations of sources of students' difficulty in mathematics and students' opportunities to learn. *Elementary School Journal, 117*(3), 345–370. https://doi.org/10.1086/690113

Zahner, W., Velazquez, G., Moschkovich, J., Vahey, P., & Lara-Meloy, T. (2012). Mathematics teaching practices with technology that support conceptual understanding for Latino/a students. *The Journal of Mathematical Behavior, 31*(4), 431–446. https://doi.org/10.1016/j.jmathb.2012.06.002