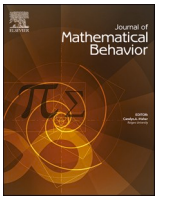
Journal of Mathematical Behavior 70 (2023) 101054

Contents lists available at ScienceDirect

Journal of Mathematical Behavior

journal homepage: www.elsevier.com/locate/jmathb



Measuring interaction quality in mathematics instruction: How differences in operationalizations matter methodologically

Kim Quabeck [a], Kirstin Erath [b], Susanne Prediger [a,c,*]

[a] TU Dortmund University, Germany
[b] Martin Luther University Halle-Wittenberg, Germany
[c] IPN Leibniz Institute for Science and Mathematics Education, Berlin, Germany

ARTICLE INFO

Keywords:
Interaction
Space for talk
Mathematical richness
Discursive richness
Video study
Operationalizations

ABSTRACT

Quality of interaction can enhance or constrain students' mathematical learning opportunities. However, quantitative video studies have measured the quality of interaction with very heterogeneous conceptualizations and operationalizations. This project sought to disentangle typical methodological choices to assess interaction quality in six quality dimensions, each of them in task-based, move-based, and practice-based operationalizations. The empirical part of the study compared different conceptualizations with their corresponding operationalizations and used them to code video data from middle school students (n = 210) organized into 49 small groups who worked on the same curriculum materials. The analysis revealed that different conceptualizations and operationalizations led to substantially different findings, so their distinction turned out to be of high methodological relevance. These results highlight the importance of making methodological choices explicit and call for a stronger academic discourse on how to conceptualize and operationalize interaction quality in video studies.

1. Introduction: Heterogeneous ways to capture interaction quality

Quantitative and qualitative research has shown that quality of instruction substantially influences what students can learn in particular classrooms (Brophy, 2000; Cai et al., 2020). Nevertheless, the authors of the international TALIS study emphasized that " we are only beginning to understand what makes a difference in terms of quality teaching" (OECD, 2020, p. 14) and called for further striving for depth in the ways instructional quality is measured quantitatively.

Whereas early quantitative coding protocols mainly captured surface structures of instruction such as activity structures (group work, seat work, and whole-class discussion; see Stigler et al., (1999) TIMSS video study) or teachers' and students' talk time (Flanders, 1970), later coding protocols were successively extended to also capture deeper structures of instruction (Bostic et al., 2021; Praetorius & Charalambous, 2018). According to the survey by Hiebert and Grouws (2007), these deeper structures are shaped by various areas of instruction such as (a) emphasis on and time allocated for different learning goals and topics, (b) kinds of tasks and representations, (c) kinds of questions and accepted responses, and the nature of discussions. In this paper, we focus on measuring the quality of interaction, which is one particular area among areas (a)-(c), and others, for example, learning objectives, assessment methods, and activity structures.

Classroom interaction has been defined as the ways in which teachers and students interact in the negotiations of meanings of

* Corresponding author at: TU Dortmund University, Germany.
E-mail address: prediger@dzlm.de (S. Prediger).

https://doi.org/10.1016/j.jmathb.2023.101054
Received 28 October 2022; Received in revised form 3 April 2023; Accepted 3 April 2023
Available online 29 April 2023

mathematical ideas (Bauersfeld, 1988). Many qualitative case studies on classroom interaction have unpacked how students learn mathematics by participating in groups' processes of negotiating mathematical ideas as they are realized by communication (Bauersfeld, 1988; Walshaw & Anthony, 2008).

When turning from in-depth qualitative research on interaction to quantitative research approaches for large video studies, the depth of analysis must necessarily be reduced (Pauli & Reusser, 2015). In various quantitative coding protocols, this has been realized in very heterogeneous ways (Bostic et al., 2021; Cai et al., 2020; Mu et al., 2022; Praetorius & Charalambous, 2018). Various researchers have therefore raised concerns that this heterogeneity and also the missing transparency might impede the comparative interpretation of results of large video studies with respect to interaction quality (Cai et al., 2020; Praetorius & Charalambous, 2018; Pauli & Reusser, 2015). Although first examples have been given that different ways of measuring quality may not necessarily capture the same phenomena (Ing & Webb, 2012; Charalambous & Praetorius, 2018), no framework yet exists for articulating the differences in systematic ways.

Therefore, this paper disentangles these (often implicit) heterogeneous conceptualizations ("what" to measure; Hiebert & Grouws, 2007, p. 376; Praetorius & Charalambous, 2018, p. 538) and operationalizations ("how" to measure; Hiebert & Grouws, 2007, p. 376; Praetorius & Charalambous, 2018, p. 539) of interaction quality. The intent of this paper is to give a theoretical contribution and an empirical contribution to the highly y methodological discourse: We contribute theoretical constructs to systematically articulate differences in the conceptualizations and operationalizations that have so far only been discussed for some examples. On this basis, the empirical study pursues the following overall research question: To what extent do different conceptualizations and operationalizations of interaction quality capture the same or different phenomena? Even if the study's findings might not be astonishing, they reveal evidence supporting the concept that the theoretical distinctions are of methodological relevance and are also worth being further considered for methodological decisions in other coding protocols.

2. Theoretical disentanglement of different conceptualizations and operationalizations of interaction quality

In this theory section, we suggest constructs for articulating differences in conceptualizations and operationalizations of interaction quality (being one out of several areas of instruction; Hiebert & Grouws, 2007). Drawing upon existing research on interaction, potential conceptualizations of what to measure in interaction quality are systematized in Section 2.1, disentangled according to (a) the quality domains (including only the pure quantity of mathematical talk or aspects of mathematical or discursive richness of the talk) and (b) their focus on teachers' intended activation, teachers' enacted activation, or students' participation. The brief literature review in Section 2.2 aims to show that all theoretically derived conceptualizations are addressed in some existing coding protocols, but with very different operationalizations, which we distinguish according to whether they are based in the tasks, the teacher moves, or collectively enacted discourse practices.

2.1. Conceptualizations of interaction quality

2.1.1. Theoretical underpinnings in qualitative case studies

Since the seminal early interactionist work on mathematics classroom interaction (Bauersfeld, 1988), many qualitative case studies have outlined the significance of high-quality or low-quality interaction as key environments that offer or constrain students' mathematical learning opportunities (Lampert & Cobb, 2003). Over three decades, a substantial body of qualitative case studies on mathematics classroom interaction have developed strong theoretical underpinnings for this significance and have provided deep insights into the interactive mechanisms in the co-construction of ideas and meanings, as shown by overviews in Lampert and Cobb (2003) and Walshaw and Anthony (2008).

Based on these understandings of the complex mechanisms of interactions, the case studies also contributed to evaluating interaction quality and identified three main quality domains that enhance students' mathematical learning opportunities:

- *Space for talk*. Qualitative researchers emphasized that students need space for talk to actively engage with mathematics (Lampert & Cobb, 2003). As not every kind of student talk is expected to be equally productive for mathematics learning (O'Connor et al., 2017), space for talk is only a prerequisite, as the richness of the talk is decisive.
- *Mathematical richness*. Various studies have outlined that students' have stronger learning opportunities when the interaction is centered around rich mathematical ideas and shaped by high cognitive demands (Walshaw & Anthony, 2008). In this context, mathematical richness involves a focus on conceptual understanding, problem-solving, or high cognitive demands posed by the tasks and maintained in the interaction (Henningsen & Stein, 1997; Hiebert & Grouws, 2007; Schoenfeld, 2014).
- *Discursive richness*. Many qualitative studies have investigated the alignment of mathematical richness with the discursive richness of interaction, which can be characterized by "the respectful exchange of ideas, ... a sustained press for justification and explanation, … orchestrate discussion and argumentation" (Walshaw & Anthony, 2008, p. 540). Discursive richness has also been described as exploratory talk (Mercer et al., 1999) and as being substantiated by the discourse practices in which students engage, which involve not only reporting procedures but also, in particular, explaining meanings and arguing about ideas (e.g., Barwell, 2012; Erath et al., 2018; Moschkovich, 2015).

Overall, qualitative in-depth case studies have revealed a high complexity of interactions and the mutual intertwinement of mathematical richness and discursive richness, but the findings have stemmed from only small samples and short extracts of interaction.

2.1.2. Quality domains and foci in conceptualizations for quantitative coding protocols

In contrast to in-depth case studies of small extracts of interaction, quantitative studies on instructional quality have aimed at covering larger samples and longer extracts, so standardized coding procedures have been developed to capture instructional quality in large video data sets, including the particular area in view of this paper: interaction quality.

One of the first quality domains being coded was students' space for talk: already Flanders (1970) showed that teachers talked 2/3 of the time in whole-class discussions. More recent studies have continued to capture students' amount of talk and have revealed a large variance between classes and between individual students within classes (e.g., TIMMS video study, Hiebert et al., 2003; PYTHAGORAS study, Pauli & Lipowsky, 2007; Sedova et al., 2019). Repeated quantitative evidence has validated the hypothesis generated in case studies that talk time alone is not predictive for measurable learning gains: neither the average student talk time in a class nor the individual student talk time (Pauli & Lipowsky, 2007, p. 110; similarly in Inagaki et al., 1998). Later coding protocols therefore increasingly tried to also capture the quality domains mathematical richness or discursive richness of interactions, with different methodological decisions on what exactly to focus on.

When quantitatively measuring mathematical or discursive richness of interactions, many coding protocol studies have been problematized to capture teachers' actions for activating students, while neglecting students' real participation beyond talk time (Charalambous & Litke, 2018; Erath et al., 2021; Praetorius & Charalambous, 2018), although these are structurally different phenomena of supply and use (Brühwiler & Blatchford, 2011; Helmke, 2009). For example, some coding protocols have measured mathematical richness almost entirely using teachers' intended activation, in other words, the demands of applied tasks or teacher moves, without taking into account the enacted demand, which means whether those intended learning opportunities were actually realized (e.g., TEDS-instruct, Schlesinger et al., 2018; review by Spreizer et al., 2022). In contrast, if the theoretical understanding is that interaction is interactively co-constructed by teacher and students (see the last section), the focus should also be on the teachers' enacted activation in the interaction, meaning the extent to which students' engagement corresponds to the intended demands. This was realized, for example, in Schoenfeld's (2014) TRU scheme, which captured mathematical richness in the teachers' offers and cognitive demand in the interplay of teachers' prompts and students' contributions.

The teachers' enacted activation mostly involves only class engagement, but the average student participation can substantially deviate from each student's individual participation (Ing & Webb, 2012). Praetorius and Charalambous (2018) criticized many papers as having not clearly distinguished class participation and individual participation. We therefore differentiated three foci of interaction quality: teachers' intended activation, teachers' enacted activation in the interaction, and individual students' participation (columns of Table 1).

These three foci of interaction quality and the introduced three quality domains of space for student talk, mathematical richness, and discursive richness served as the theoretical constructs which allowed us to derive nine possible conceptualizations of interaction quality from the literature. These nine conceptualizations are systematized as entries in Table 1 spanned by quality domains for interaction (in rows) and foci of interaction quality (in columns).

The quality domains do not capture totally different aspects of interaction quality but overlap as they all refer to mathematical talk. We decided to include space for talk as a starting point of classroom talk in our conceptualization as no space for talk involves no discursively or mathematically rich talk, but not all mathematical talk is equally important for learning (Howe et al., 2019). So far, the association between all mathematical talk (independent of richness) and mathematically and discursively rich talk has not gotten much attention (one exception was provided by Sedova et al., 2019). By including the space for talk, we can empirically investigate this relationship. Further, even though qualitative case studies have given many examples of overlaps of mathematically and discursively rich talk (e.g., Erath et al., 2018; Moschkovich, 1999; Walshaw & Anthony, 2008), their relation has rarely been quantified. A first quantitative study of their overlap indicated that they correlate, but not with very high correlations: Prediger and Neugebauer (2021) reported correlations of discursive demands with mathematical richness of 0.50 and with cognitive demands of 0.40 (with holistic operationalizations of these quality domains without distinguishing the teachers' supply and the students' use). This suggests the need to distinguish the quality domains analytically in order to quantify and scrutinize the expected overlap.

In the following section, we will show that indeed, all these theoretically possible conceptualizations have been used in some existing coding protocols, with different names standing for the same conceptualizations or the same names for different conceptualizations. So the introduced constructs of quality domain and focus might help to substantiate a systematic methodological discourse.

Table 1
Different conceptualizations of interaction quality by a combination of foci (in columns) and quality domains for interaction quality (in rows).

Quality domains	Teachers' intended activation	Teachers' enacted activation in the interaction	Individual students' participation
Space for student talk	Offered space for student talk (e.g., by teachers' questions)	Class engagement in student talk	Individual participation in student talk
Mathematical richness	Conceptual and other high cognitive demands and supports (e.g., by tasks, representations, and teachers' moves)	Class engagement in rich mathematical (e.g., conceptual instead of procedural) activities	Individual participation in rich mathematical 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)

2.2. Systematizing heterogeneous operationalizations for interaction quality in existing coding protocols

Given the heterogeneity of existing coding protocols, Praetorius and Charalambous (2018) and Mu et al. (2022) called for a more explicit methodological discourse on how to operationalize the chosen conceptualizations for manageable analyses of larger video data sets. In order to facilitate this methodological discourse, we suggest some constructs for distinguishing operationalizations so that they can be articulated, compared, and discussed (Ing & Webb, 2012; Praetorius & Charalambous, 2018) and suggest a systematization of methodological decisions taken in existing coding protocols for interaction quality.

For this systematization, we reviewed a (necessarily incomplete) selection of mathematics-specific coding protocols with the aim of identifying the conceptualizations and operationalizations underlying the studied quality features. We started with the first comprehensive international quality of teaching study, the TIMSS video study (Stigler et al., 1999). Then we added more recent studies on quality assessments which were evaluated as valid in the review by Bostic et al. (2021): IQA (the lesson observation part in Boston, 2012), EQUIP (instructional and discourse factors in Marshall et al., 2010), MQI (Charalambous & Litke, 2018), RTOP (Sawada et al., 2002), TRUMath (Schoenfeld, 2013), and OTOP (Flick et al., 2004). As this set of reviewed coding protocols predominantly focused on

Table 2
Quality features identified in video studies in three domains for interaction quality with operationalization base.

Quality domain	Focus	Quality feature	Coding protocol in which quality feature occurs	Operationalization base
Space for student talk	Intended activation	• Teacher's questions	• TIMSS	• M, P
	Enacted activation	• Communicative interactions (several features)	• RTOP	• P
		• Students' utterances	• PYTHAGORAS	• P
	Individual participation	• Students' utterances	• IQA	• P
			• Sedova et al. (2019)	• P
Mathematical richness	Intended activation	• Potential of tasks	• IQA	• T
		• Intention of moves for cognitive activation	• PYTHAGORAS	• M
		• Potential for cognitive activation (modeling and argumentation)	• COACTIV	• T
	Enacted activation	• Task implementation	• IQA	• T, P
		• Cognitive demand	• IQA	• T, M, P
		• Teachers'/students' linking	• IQA	• T, M, P
		• Order of instruction	• EQUIP	• P
		• Knowledge acquisition	• EQUIP	• P
		• Student role	• EQUIP	• T, M, P
		• Content depth	• EQUIP	• P
		• Conceptual development	• EQUIP	• T, M, P
		• Richness of the mathematics	• MQI	• T, M, P
		• Errors and imprecisions	• MQI	• T, M, P
		• Students' practices	• MQI	• P
		• Propositional knowledge	• RTOP	• T, M, P
		• Coherence and accuracy	• TRUMath	• T, M, P
		• Cognitive demand, use of assessment	• TRUMath	• M, P
		• Students' problem solving/metacognition	• OTOP	• M, P
		• Challenge of ideas/multiple responses	• OTOP	• M, P
		• Conceptual thinking/connections/multiple representations	• OTOP	• T, M, P
	Individual participation	• (To our knowledge not captured quantitatively for individual students)		
Discursive richness	Intended activation	• Teacher's questioning demand	• TALIS	• M
			• TIMSS	• M
	Enacted activation	• Teacher's press	• IQA	• M, P
		• Students' providing	• IQA	• P
		• Questioning level/complexity/ecology	• EQUIP	• M
		• Communication pattern/classroom interactions	• EQUIP	• M
			• TRUMath	• M, P
		• Agency: authority and accountability	• OTOP	• P
		• Students' discourse		
	Individual participation	• Students' reasoning	• Sedova et al. (2019)	• P
		• Number of explanations per problem	• Ing & Webb (2012)	• P
		• Students' contributions	• PYTHAGORAS	• P

Note. T = task-based, M = move-based, P = practice-based

teachers' activation rather than students' individual participation and rarely on discursive richness, we enriched our selection with the following widely cited coding protocols to compensate for the one-sidedness: PYTHAGORAS (Lipowsky et al., 2007; Hugener et al., 2006; Pauli & Lipowsky, 2007), Ing & Webb (2012), COACTIV (Neubrand et al., 2013; Kunter et al., 2013), Sedova et al. (2019), and TALIS (OECD, 2020).

In Table 2, the nine conceptualizations from Table 1 are organized in the first two columns. The third column of Table 2 contains the identified quality features from the analyzed coding protocols. They are arranged according to their corresponding conceptualization for interaction quality, for example, the quality feature cognitive demand from IQA to the quality domain mathematical richness and the focus enacted interaction. Afterwards, also in reference to the coding manuals, we analyzed the applied operationalization bases that were decisive for rating high or low quality. The assignment was then made according to the intention of the coding protocol. For example, if the mathematical richness was more central than the discursive richness in the theoretical foundation or aims of the coding protocol, the quality feature was assigned to mathematical richness. As this process partly involved decisions relying on interpretations, decisions were reached through consensus between the authors and further researchers from the Dortmund research group.

The overview in Table 2 shows that for each of the nine theoretically derived conceptualizations, diverse quality features have been studied. Sometimes they have almost the same names (e.g., students' utterances in PYTHAGORAS and students' providing in IQA), but this does not mean they were intended to measure the same conceptualization (as problematized by Mu et al., 2022; Praetorius & Charalambous, 2018). Yet there is an unequal occurrence of the conceptualizations: Most of the quality features identified in existing coding protocols have referred to the focus enacted activation, whereas less often, teachers' intended activation and student individual participation have been focused on. Furthermore, more quality features have been studied for the domain mathematical richness than for discursive richness or space for students' talk.

The last column of Table 2 denotes how the quality features are operationalized, and for this we distinguished three different bases for how to capture mathematical or discursive richness:

- Task-based operationalizations are used for rating the quality of demands and supports posed by tasks and representations. For example, a task can request students to practice a calculation procedure that has a lower demand than requesting an explanation for why a specific procedure works for particular numbers.
- Move-based operationalizations are used for rating the quality of demands and supports posed by teachers' moves in the classrooms, for instance, posed questions. Move-based operationalizations refer to the teacher's intention for the interaction; thus, the focus is more on the teacher, in contrast to the jointly enacted practices that are focused on in practice-based operationalizations.
- Practice-based operationalizations are used for rating the quality of interactively, jointly established, collectively, or individually enacted discourse practices, for example, joint explanations. Both the teacher's and the students' actions and utterances and their interplay are taken into account.

Within the same conceptualization, the lines of Table 2 reveal a large heterogeneity in how the same quality dimensions (e.g., enacted discursive activation) are operationalized. Interaction quality is assessed based on tasks, moves, practices, or sometimes a mixture of these operationalization bases, often without clarity as to exactly which of the three is decisive for rating higher or lower quality. Table 2 does not document the multiple further measurement decisions taken for operationalizing the quality features, such as a rating of fixed or flexible time segments (from 5 min up to whole lessons) or rating of tasks; word-, sentence-, or turn-related frequencies; or specific shares (e.g., numbers of explanations per problem in Ing & Webb, 2012).

In the following, we discuss the overview for each of the quality domains: The quality domain space for student talk was operationalized mainly in practice-based ways, either in the enacted activation measuring all students' utterances (RTOP; Sawada et al., 2002) or in individual students' participation in measuring individual students' utterances (IQA; Boston, 2012). In the early TIMSS video study (Stigler et al., 1999), space for talk was operationalized by teachers' moves, so only in the intended activation. This shows the heterogeneity of operationalizations even for the allegedly easiest quality domain, space for student talk. Also, the measurement decisions varied: Whereas some studies conducted a rating of rather large time segments (5-point scale in RTOP) other studies used turn- or time-related frequencies.

The conceptualizations of the quality domain mathematical richness comprise several aspects, such as problem solving, proofing, and arguing. Hiebert and Grouws (2007) further found that mathematically rich learning opportunities tend to be shaped by teachers addressing and supporting students' conceptual understanding (as opposed to a mere focus on procedures). Due to the multiple aspects constituting mathematical richness, it is challenging to compare different studies. In addition, the quality features are named differently and entail often slightly different conceptualizations (e.g., richness of the mathematics, MQI; conceptual development, EQUIP; or conceptual thinking, OTOP). Concerning the foci within mathematical richness, researchers have often focused on the conceptual richness of teachers' intended activation when capturing only the task demands but not their enactment in classrooms (e.g., potential of the task in IQA; Boston, 2012) or teachers' enacted activation in the ways teachers engage their classes in conceptual activities. This class engagement in conceptual activities has been captured in diverse operationalizations, by rating the quality of the task implementation by teaching practice (IQA; Boston, 2012), by the enacted move demand (Hugener et al., 2006), or in the conceptual practices that are co-constructed in the interaction (e.g., in some 4-point scale ratings in MQI; Charalambous & Litke, 2018). Even the conceptual practices co-constructed in the interaction have been operationalized diversely and have differed, for example, in the extent to which they mainly comprise teachers' enactment of mathematical richness in interaction (e.g., depths of the mathematics offered) or focus more on class engagement (e.g., level of student work). However, the published coding manuals have often not explicated the exact operationalization bases (task, moves, or practices) for the quality assessment. Often, several bases are mentioned

and combined, but without explicating the exact relation between the combined bases.

Capturing the quality domain discursive richness has been challenging due to its interactive constitution in classroom talk. Thus it is often simplified in quantitative research approaches (Pauli & Reusser, 2015), for example, when researchers use a shortened period for analyzing discursive richness (e.g., 30 utterances in TIMSS; Stigler et al., 1999). Some studies have captured teachers' intended activation using move-based operationalizations (e.g., TIMSS; Stigler et al., 1999) rather than enacted classroom discourse practice. Other studies have utilized a combination of moves and practices (e.g., IQA, Boston, 2012), or have focused mainly on practices (e.g., students' discourse in OTOP; Flick et al., 2004) to capture the enacted discursive richness in the interaction. An individual student's participation has been assessed by counting the number of students' utterances with reasoning (e.g., Sedova et al., 2019) or by counting the contribution lengths in word or time-related measurements (e.g., PYTHAGORAS; Pauli & Lipowsky, 2007). This implies a considerable simplification of the individual student's discursive activation or participation, which are criticized as under-complex approaches (Pauli & Reusser, 2015), rated only roughly in larger time segments or by counting the length of students' sentences without taking into account different degrees of richness in the discourse practices. Summing up this brief literature review, we have seen a large variety of operationalizations applied in existing coding protocols for quantitatively capturing interaction quality in mathematics classrooms. The studied quality features have differed, particularly in terms of their focus on teachers' intended activation, teachers' enacted activation in the class engagement, and individual students' participation (see Table 2). Furthermore, researchers have utilized task-based, move-based, or practice-based operationalizations for judging the quality. In addition, different measurement decisions have been used that were likely to influence the studies' empirical results (Ing & Webb, 2012), often without a well-substantiated methodological discourse on these differences and methodological choices.

2.3. Refined research question

The heterogeneity of conceptualizations and operationalizations for interaction quality raises the methodological question of how far these methodological decisions influence what empirical phenomena are captured (Cai et al., 2020; Mu et al., 2022; Praetorius & Charalambous, 2018).

Differences between teachers' intended activation and teachers' enacted activation have often been shown (Columns 2 and 3 in Table 1; e.g., Hiebert et al., 2003; Cai et al., 2020). But the difference in quality features of interaction focusing on teachers' enacted activation in the whole group of students and the participation of individual students in different operationalizations (Columns 3 and 4 in Table 1) have not yet received much interest. The same applies to the relation between discursively rich student talk and non-qualified student talk, which has not yet been well investigated quantitatively despite all the qualitative findings emphasizing the difference.

There have been studies comparing quality ratings with different coding protocols, in other words, focusing on the empirical comparison of quality ratings when different conceptualizations are used (e.g., Brunner, 2018). However, there has been less interest in empirically examining different operationalizations for the same conceptualizations than in comparing different conceptualizations (Ing & Webb, 2012). Thus, the scientific interest generally lies in investigating the extent to which it is worthwhile to decompose operationalizations in more detail.

One reason why the subtle differences might not matter in many studies is that the chosen learning goals and selected tasks and representations already heavily influence the interaction quality. So studying the subtle differences captured with different conceptualizations and operationalizations for interaction quality is often confounded with these areas of instruction. That is why we decided to compare different conceptualizations and operationalizations in a video data corpus in which three other areas of instruction (listed by Hiebert & Grouws, 2007, p. 379) are held constant—the learning goals, the time allocated to particular topics, and the kinds of tasks and representations—while most of the interactions between teachers and students differ. On this background, we refine our research question as follows:

How do different conceptualizations of interaction quality with their different operationalizations—task-based, move-based, and practice-based—correlate in a given video data corpus where the learning goals, tasks, and representations remain constant?

3. Methodological framework for the empirical comparison of conceptualizations and operationalizations

3.1. Research context and data corpus

To pursue our research question, we draw upon a video data corpus gathered in the study MuM-Mesut (Prediger et al., 2022). Videos were recorded during small-group instruction that was aimed at developing conceptual understanding of fractions and their operations, with an integrated focus on enhancing students' meaning-related language. The small-group instruction followed four design principles: engaging students in rich discourse practices (DP1), connecting multiple languages and representations (DP2), using macro-scaffolding in combined language- and content-learning trajectories (DP3), and building up an integrated shared meaning-related vocabulary before formal vocabulary (DP4; Prediger et al., 2022; Wessel & Erath, 2018). Students who participated in the intervention showed significantly higher learning gains than a control group with business-as-usual teaching (Prediger et al., 2022).

The video-recorded instruction was organized in teacher-led small groups (3–6 students) and spanned over five sessions of 90 min each. The groups were taught by pre-service teachers in their master's studies or PhD students with teaching certificates. All lessons were video-recorded and students' working materials were collected. Because the aim was to investigate the interdependencies between quality features and their (task-based, move-based, and practice-based) operationalizations, all students worked on nearly

identical tasks with the same representations. Some small groups received additional integrated lexical support (Wessel & Erath, 2018). By keeping curriculum resources nearly identical and standardizing teachers' intended activation in collective preparation sessions, differences in teachers' enacted activations in the interactions and individual student's participation became observable in a fine-grained way.

The paper's empirical section draws on data from 210 middle school students aged 10–14 years old participating in 49 small groups. About 30 h of video were analyzed that covered small groups' work on knowledge inquiry, consolidation, and organization for reinventing the concept of equivalence of fractions.

3.2. Analytic framework for the presented empirical study

To study the commonalities and differences of various operationalizations, we created an analytic framework covering different conceptualizations from the theoretically derived cells in Table 1. To reduce complexity, we restricted the mathematical richness to conceptual richness (as opposed to a procedural focus; Hiebert & Grouws, 2007), which is included in all coding protocols (see Table 2). Given that intended and enacted activation has already been compared in other studies (Stigler et al., 1999), we did not focus on teachers' intended activation but rather further pursued the two foci related to enacted activation and individual participation that sometimes occur as mixed in the existing coding protocols (see Table 2). Thus, each of the three quality domains was split into an activation dimension (conceptualized as class engagement) and a participation dimension, resulting in six refined conceptualizations (Fig. 1).

For the operationalization of the refined conceptualizations into operationalized quality features (see Fig. 1), we systematically distinguished task-based, move-based, and practice-based operationalizations as derived and harmonized from existing coding protocols. This distinction will later allow us to study their relationship empirically and validate or invalidate implicit assumptions about their associations.

Within the two quality dimensions of space for student talk, no qualification of richness is necessary, so space for student talk is simply operationalized as the relative length of all students' talk (talk-related activation, abbreviated TA) or individual students' talk (talk-related participation, abbreviated TP) in relation to the whole time on task. This is similar to Flanders (1970) and also to more recent coding protocols such as RTOP or IQA.

In contrast, the other four quality dimensions require further operationalizations of conceptual richness and discursive richness. The literature review in Table 2 revealed that the richness is often captured in holistic ratings of larger time segments (Praetorius & Charalambous, 2018), but these holistic ratings in coding protocols have been criticized due to the quick loss of information (Pauli & Reusser, 2015) with their simplified measurement. It has been shown that video-lite coding or not tracking individual students' actions can lead to more divergent empirical results than when operationalizing them with more detail (Ing & Webb, 2012). To substantiate the methodological choices in video-lite codings and holistic ratings and to ensure comparability with the talk-related dimensions, we chose a video-intensive coding by scrutinizing time measurements of interaction periods that meet a specific quality bar for the tasks, moves, or practices.

To capture the conceptual or discursive richness of teachers' and students' engagement in task-based, move-based, and practice-based operationalizations, we developed basic codings for identifying the conceptually or discursively rich tasks (similar to Jordan et al., 2008), for identifying the conceptually or discursively rich teachers' moves (distinction according to Hiebert & Grouws, 2007, for conceptual richness and Wessel & Erath, 2018, for discursive richness), and for identifying the conceptually and discursively rich discourse practices interactively established in the interaction (as formerly only conducted in qualitative studies, e.g., Erath et al., 2018). This allows operationalization of conceptual and discursive richness in activation and participation by the relative length of time spent in conceptually and discursively rich interaction, characterized by the identified conceptually or discursively rich tasks, moves, or practices. The details of the analytic framework are presented in the next section.

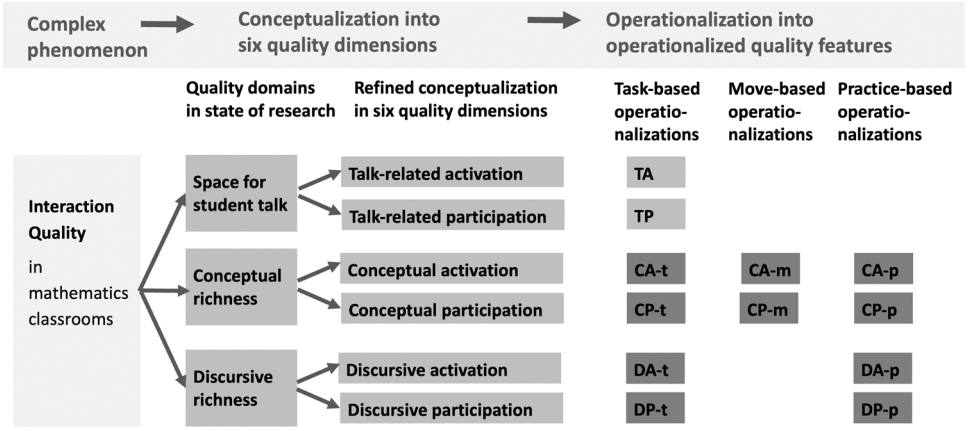


Fig. 1. Successively refining the assessment of interaction quality in the analytic framework of this paper.

3.3. Methods of data analysis: quality features with different operationalizations

The analytic framework was realized and applied in a data analysis consisting of four steps:

3.3.1. Step 1: Basic coding

The following basic codings were conducted on video data, transcripts, or curriculum resources from all 49 small groups in Transana (Multiuser 3.32d). Student assistants were introduced to the low-inferent basic coding and intensively trained for the high-inferent basic coding based on a coding manual:

- The basic coding of time on task was captured by measuring the group time spent on selected tasks, excluding breaks, organization, and off-topic time. Capturing the total time on task serves to calculate all small groups' relative lengths (for comparable values of the different operationalizations between 0 and 1). In the basic coding utterances, the length of each student's and teacher's utterances was measured in seconds. The length of utterances was used in the high-inferent codings for further quality coding steps.
- For the basic coding of task demand, all 35 tasks with 78 subtasks were coded from two perspectives. Similar to Jordan et al. (2008), tasks were categorized according to their focus on lexical, procedural, or conceptual knowledge for grasping the mathematical demand. The quality bar of conceptually rich tasks was met by tasks with high conceptual demands. Following Erath et al. (2021), the tasks were identified as discursively rich when entailing discursive demands of explaining or reporting procedure, or no discursive demand (such as stating a pure number).
- For the two high-inferent basic codings (teachers' move demand and discourse practices), a preparatory step in data analysis was necessary: identifying sense-making units. They begin with teachers' or students' initiation of a problem or question and end when the problem or question is interactively marked as being finished (similar in Schwarz et al., 2021). In cases of non-conformity between coders, a sense-making unit's length was agreed upon in a consensus-building process. The discrepancies were recorded and discussed regularly in the team of all coders. To identify conceptually rich moves, conceptual or procedural knowledge demands or neither of these were coded. The basic coding of discourse practices comprises the teachers' and students' talk in discourse practices, distinguishing non-discursive practices (such as one-word answers and less than half sentences) and rich discourse practices (such as reporting, explaining, and arguing), with or without conceptual focus. The predominant discourse practice of each sense-making unit was assigned to the whole unit. In this way, the students' and teachers' talking length (in seconds) in the respective discourse practice (sequential, integrated, or non-discursive) can be measured. Illustrative examples of discourse practices that have been identified as discursively rich or not have been documented in Wessel and Erath (2018) and Quabeck and Erath (2022).

3.3.2. Step 2: Control of interrater reliability

The two high-inferent basic codings were coded independently by two raters. Interrater reliability was controlled for the high-inferent codings in R Studio (version 3.6.3, DescTools package). The determined Cohen's κ of 0.90 for teachers' move demands and 0.91 for the richness of discourse practices indicated that interrater reliability was very good (Döring & Bortz, 2016) for both high-inferent basic codings.

3.3.3. Step 3: Deriving quality features from the basic coding

Table 3 shows how all the quality features are operationalized. In order to ensure comparability, all quality features are determined as the relative length of group time or the relative length of group talk spent in each feature, given in percent out of total time on task. The limitations of this time-related measurement are discussed in Section 5.2. Based on the basic codings, conceptual and discursive richness in activation and participation are operationalized by the relative length of time spent in conceptually and discursively rich interaction, characterized by the identified conceptually or discursively rich tasks, moves, or practices. So, the quality features are measured in proportions between 0 % and 100 %.

Table 3
Analytic framework for quality features for interaction operationalized in task-, move-, and practice-based ways (Relative length of time spent in each feature, 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 conceptual tasks	CA-m Relative length of group time spent on conceptual moves	CA-p Relative length of group talk spent on conceptual practices (including teacher)
Conceptual participation	CP-t Relative length of individual talk spent on conceptual tasks	CP-m Relative length of individual talk spent on conceptual moves	CP-p Relative length of individual talk spent on conceptual practices
Discursive activation	DA-t Relative length of group time spent on discursive tasks		DA-p Relative length of group talk spent on rich discourse practices (including teacher)
Discursive participation	DP-t Relative length of individual talk spent on discursive tasks		DP-p Relative length of individual talk spent on rich discourse practices

For example, the quality of conceptual activation is operationalized by the relative length of group time spent on conceptually rich tasks and moves (CA-t, CA-m) and by the relative length of group talk on conceptually rich practices (CA-p). A high proportion of CA-t means, that the teacher focuses the discussion more often on conceptual tasks than on procedural tasks. In groups with a high proportion of CA-m, the teachers engage the students for a relatively long time with their conceptually rich moves, for instance, by demanding the connections of representations for the part-whole concept.

A high proportion of CA-p means that the group engages for a comparatively greater length of time in the enactment of verbal conceptual practices such as explaining the meaning of a mathematical concept. For conceptual participation, we refer again to the relative length of talk time spent by individual students in contributing to conceptually rich tasks (CP-t), conceptually rich moves (CP-m), and conceptual practices (CP-p), expressed as a percent of the total time on task.

The discursive activation is operationalized by newly developed high-inferent basic codings (Table 4) but is similar to the conceptual activation as the relative length of group time spent on the identified discursively rich tasks (DA-t), for instance, tasks demanding oral or written explanations, justifications, or reasoning, and the group talk on discourse practices (DA-p), for instance, interactively negotiating mathematical meanings or reporting procedures. The discursive activation is not operationalized by teachers' moves (no DA-m), because the same articulated teacher move might initiate students explaining meanings or telling stories, depending on the norms established in the classes. Therefore, moves are not a valid base for operationalizing discursive richness. For discursive participation, we apply two quality features that comprise the relative length of talk time spent by individual students in contributing to discursive tasks (DP-t) and discursive practices (DP-p).

3.3.4. Step 4: Calculating statistics

Descriptive statistics were conducted in R studio (package DescTools) for determining the mean (as an average relative length in relation to time on task across all 49 groups), standard deviation, and Pearson's product moment correlation r between 13 out of 14 quality features. Due to technical challenges, the length of individual talk in conceptual move demands was only captured in a subsample of 29 groups, so the quality feature CP-m is only reported in the descriptive statistics but excluded from investigating correlations. Additionally, a linear regression was conducted for investigating the association between three quality features with strong overlap.

4. Quantitative results on connections between quality features

To analyze the extent to which the quality features for interaction with their task-based, move-based, and practice-based operationalizations are associated in our video data corpus, we first present all descriptive results for the captured quality features (Section 4.1) and then the correlations between quality features (Section 4.2).

4.1. Overview of descriptive results for different conceptualizations and operationalizations

Within the analytic framework, interaction quality was conceptualized in six quality dimensions, each of them operationalized in different ways into 14 quality features (Fig. 1). Table 4 documents the descriptive results: the mean (as an average relative length in relation to time on task across all 49 groups) and standard deviation for each quality feature, as recommended by Döring and Bortz (2016), for transparency.

The mean of talk-related activation (33.4 % for TA) corresponds to Flanders' (1970) classical "one third" for all students' talk time. The mean of individual students' talk-related participation (7.7 % for TP) was high, compared to whole-class settings (e.g., in TIMSS 1999, Hiebert et al., 2003), as individual students tend to have more space for talk in small-group settings. However, even these most superficial quality features had very high standard deviations (SD 13.7 % for TA and SD 5.2 % for TP), which means that the class

Table 4
Distribution of quality features with mean (M), and SD for relative lengths in the 49 groups, (CP-m only in a subsample of 29 groups, therefore excluded in the following analyses).

Quality dimension	Quality feature	Operationalization of the quality features (Relative length of time spent in each quality, given in percent in relation to total time on task of the group)	m (SD) for relative length
Talk-related activation	TA	Relative length of all students' talk	33.4 % (13.7 %)
Talk-related participation	TP	Relative length of individual talk	7.7 % (05.2 %)
Conceptual activation	CA-t	Relative length of group time spent on conceptual tasks	77.4 % (10.9 %)
	CA-m	Relative length of group time spent on conceptual moves	35.3 % (12.9 %)
	CA-p	Relative length of group talk spent on conceptual practices	23.5 % (12.3 %)
Conceptual participation	CP-t	Relative length of individual talk spent on conceptual tasks	5.5 % (04.1 %)
	CP-m	Relative length of individual talk spent on conceptual moves	4.5 % (03.1 %)
	CP-p	Relative length of individual talk spent on conceptual practices	2.8 % (02.7 %)
Discursive activation	DA-t	Relative length of group time spent on discursive tasks	58.4 % (14.2 %)
	DA-p	Relative length of group talk spent on rich discourse practices	33.0 % (11.8 %)
Discursive participation	DP-t	Relative length of individual talk spent on discursive tasks	4.6 % (03.8 %)
	DP-d	Relative length of individual talk spent on rich discourse practices	3.9 % (03.2 %)

engagement as captured by talk-related activation TA and the individual talk-related participation TP varied strongly between the 49 small groups.

For each of the quality dimensions conceptual activation and conceptual participation, three quality features with different operationalizations can be compared. Their means differed strongly for different operationalizations: task-based operationalizations revealed substantially higher means than move-based operationalizations and practice-based operationalizations. With respect to conceptual activation, the small groups spent most of the time on conceptual task demands (77.4 % of CA-t); however, teachers did not enact all the time on task with rich conceptual moves (35.3 % for CA-m) or in rich conceptual practices (23.5 % for CA-p). For the individual conceptual participation, the relative length of individual talk on conceptual tasks (5.5 % for CP-t) exceeded the relative length of individual talk enacted after conceptual moves (4.5 % for CP-m) and the individual talk enacted in rich conceptual practices (2.8 % for CP-p), all in percent of total time on task. Thus, similar to the conceptual activation operationalized by task demand, not every student's contribution seemed to meet the demands necessary for participating in conceptual practices for constructing mathematically rich conceptual knowledge, even when articulated on conceptual tasks and after conceptual moves.

For discursive activation, the average relative length of group time spent on discursively rich tasks (58.4 % for DA-t) exceeded the relative length of group talk spent in rich discourse practices (33.0 % for DA-p). This is similar to the results in the conceptual activation; however, relative length of group time spent on discursively rich tasks was lower than the relative group time spent on conceptual tasks. The difference also existed between the individual participation when assessed with the task- or practice-based operationalizations.

4.2. Correlations between operationalizations for the same quality dimensions

4.2.1. Overview of correlations between all quality features

The differences in the distribution of quality features for the same quality dimension but with different operationalizations leads to the question of how exactly the quality features based on different conceptualizations and operationalizations are associated. In Table 5, we present all correlations between all quality features, the dependencies between differently operationalized quality features of the same quality dimensions, and quality features from different quality dimensions with the same operationalization. Overall, the range of correlations between $r = 0.01$ and 0.93 was remarkable, the significant correlations ranged from irrelevant associations of 0.15 via medium associations between 0.3 and 0.5 and strong associations above 0.5 (Cohen, 1988). Low correlations (below 0.3) and medium correlations (between 0.3 and 0.5) must be interpreted as measuring different phenomena, even for quality features previously treated as measuring the same. For example, the correlation between a task-based operationalization of conceptual activation and (CA-t) and a practice-based operationalization of conceptual activation (CA-p) was only $r = 0.15$. This means that the relative length of group time spent on conceptually rich tasks was not necessarily closely related to the intensity in which students were engaged in conceptual practices.

The five quality features capturing individual students' participation (TP, CP-t, CP-p, DP-t, and DP-p) seemed to be more closely associated to each other (r between 0.67 and 0.93) than the eight quality features capturing teacher activation (r between 0.01 and 0.81).

The connection between activation and participation within the same dimension and operationalization (e.g., DA-t and DP-t) was highly varied and worth being discussed in more depth in the next subsections.

4.2.2. Correlations between operationalized quality features in the same quality dimension

It is of high methodological importance to discuss in more depth how the quality features within the same quality dimensions differ or coincide when operationalized differently. Therefore, this section discusses the correlations between task-based, move-based, and practice-based operationalizations. Talk-related quality dimensions are not considered since they are operationalized in only one way.

Given the absence of quantitative comparisons of different operationalizations and the missing attention to possible differences, the implicit hypothesis so far might have been that the operationalizations might all capture the same phenomena within a conceptualization, so the correlations might be expected to be high, in particular for move-based and practice-based operationalizations, which

Table 5
Correlations (Pearson's r) between quality features in the video data set of 49 groups and 210 students (correlations marked in bold are significant with $p < 0.05$)

Relative length of … out of time on task	TA	TP	CA-t	CA-m	CA-p	CP-t	CP-p	DA-t	DA-p	DP-t	DP-p
TA | … all students' talk	**1**	**0.57**	**0.21**	**0.19**	**0.14**	**0.52**	**0.27**	0.13	**0.27**	**0.49**	**0.32**
TP | … student's individual talk		**1**	0.11	0.07	0.09	**0.93**	**0.67**	0.07	**0.14**	**0.86**	**0.75**
CA-t | … group time spent on conceptual tasks			**1**	**0.20**	**0.15**	**0.25**	**0.16**	**0.57**	**0.16**	**0.29**	**0.15**
CA-m | … group time spent on conceptual moves				**1**	**0.45**	0.11	**0.21**	0.03	**0.65**	0.03	**0.26**
CA-p | … group talk spent on conceptual practices (incl. teacher)					**1**	0.13	**0.46**	0.13	**0.67**	0.13	**0.26**
CP-t | … individual talk spent on conceptual tasks						**1**	**0.67**	0.07	**0.17**	**0.85**	**0.73**
CP-p | … individual talk in conceptual practices							**1**	0.09	**0.32**	**0.64**	**0.89**
DA-t | … group time spent on discursive tasks								**1**	0.01	**0.43**	0.04
DA-p| … group talk spent in rich discourse practices									**1**	0.1	**0.4**
DP-t | … individual talk spent on discursive tasks										**1**	**0.67**
DP-p | … individual talk spent on rich discourse practices											**1**

qualitative studies have often showed to have interplay (e.g., Webb et al., 2008).

These possibly hypothesized connections were only partially found in the data (Table 5). Fig. 2 depicts the selected correlations for conceptual activation and conceptual participation from Table 5 in a graphical way: The horizontal black lines show the correlations between quality features within the same quality dimension and the grey lines between the two quality dimensions.

For conceptual participation, the correlation between the task-based quality feature CP-t and the practice-based CP-p was high ($r = 0.67$) but was not 1. This means that the time spent by an individual student talking about conceptual tasks was not necessarily filled with conceptual practices, and those students that participated most in conceptual tasks tended to substantially contribute to conceptual practices, but not necessarily to most.

For conceptual activation, the practice-based operationalization correlated moderately with move-based operationalization ($r = 0.45$). This means that what teachers' moves demanded for the oral discussion in the conceptual activation was moderately associated with teachers' and students' actual conceptual practices (see Quabeck & Erath, 2022, for a qualitative illustration from a transcript). In contrast, the task-based quality feature CA-t correlated with the move-based and practice-based quality features only weakly, with r of 0.15 and 0.2. This means that more time in conceptual tasks was only loosely associated with an emphasis on conceptual moves and only loosely associated with time spent on conceptual practices.

Between conceptual activation and conceptual participation, the highest correlation was between CA-p and CP-p ($r = 0.46$), but it was lower ($r \leq 0.25$) for all other associations between activation (CA-t, CA-m, and CA-p) and individual participation (CP-t and CP-p).

Fig. 3 depicts the same information for discursive activation and discursive participation, which were only operationalized by task-based and practice-based features, not by move-based features. Again, the task-based and practice-based operationalizations of discursive participation showed high correlations ($r = 0.67$ for DP-t and DP-p). In contrast, the task-based (DA-t) and practice-based operationalizations (DA-p) of discursive activation were not connected in our data ($r = 0.01$, non-significant). Qualitative insights into the video data explain the not significant association: We found examples of teachers spending more time on tasks with low discursive demands but still eliciting rich discourse practices whereas some teachers spent a great deal of time in discursively rich tasks but without engaging the group in rich discourse practices. Thus, our data showed that working with the same discursive task demands did not automatically imply an enactment of rich discourse practices. However, this points towards more emphasis on the investigation of small groups' individual student's participation in interactions.

Between discursive activation and discursive participation, the correlations are moderate between the operationalizations within the same bases ($r = 0.40$ for DA-p and DP-p; $r = 0.43$ for DA-t and DP-t), and non-significant and very small for the two other combinations ($r \leq 0.1$). This means, for instance, that the class engagement in discursively rich tasks (DA-t) did not necessarily coincide with an individual student's rich discursive participation, assessed by the relative length of individual talk in rich discourse practices (DP-p). A student's individual talk in rich discourse practices also occurred in other task demands.

4.2.3. *Correlations between quality features of different quality dimensions and similar operationalizations*

In this section, we investigate the influence of the conceptualizations (in the vertical relations in Table 3), in other words, how talk-related, conceptual, and discursive quality dimensions were related. The previous section revealed that operationalizations mattered substantially, and activation and participation were associated only moderately, which is why only correlations between features with the same operationalization base were considered. As move-based operationalizations were only available for the conceptual dimensions, we restrained our analysis to task-based and practice-based operationalizations, each for both activation and participation.

Qualitative case studies have been interpreted to suggest that no space for talk involves no discursively or mathematically rich talk, and mathematically and discursively rich talk always coincide (e.g., Moschkovich, 1999; Walshaw & Anthony, 2008), which would suggest high correlations, or at least 0.40/0.50 as in Prediger and Neugebauer (2021). As talk is often at the same time conceptually and discursively rich, we expect a closer connection between those two quality domains than to the talk-related quality dimensions.

These hypothesized correlations were indeed found in the data set in Table 5, more for individual participation than for activation. Fig. 4 depicts the correlations of talk-related, conceptual, and discursive activation, on the left between the task-based operationalizations and on the right between the practice-based operationalizations.

The task-based operationalizations with and without qualifying the richness were not strongly associated ($r = 0.13$ for TA – DA-t, non-significant, and $r = 0.21$ for TA – CA-t), but moderately associated for task-based conceptual and discursive activation ($r = 0.57$).

For the practice-based operationalizations, the correlations between conceptualizations were moderate ($r = 0.27$ for TA – DA-p and

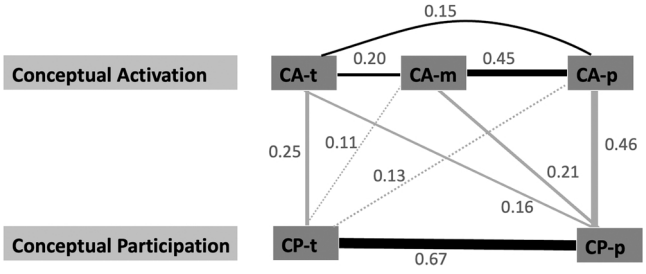


Fig. 2. Correlations between quality features in task-based, move-based, and practice-based operationalizations for conceptual activation and conceptual participation (non-significant correlations are depicted with dotted lines).

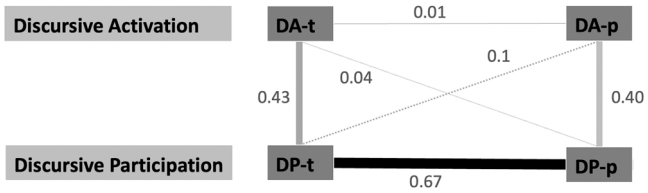


Fig. 3. Correlations between quality features in task-based, and practice-based operationalizations for discursive activation and discursive participation (non-significant correlations are depicted in dotted lines).

r = 0.67 for DA-p – CA-p) with one exception (r = 0.14 for TA – CA-p).

Comparing the correlations of the task- and practice-based operationalizations, the conceptual and discursive dimensions (DA-t – CA-t; CA-p – DA-p) were in closer connection (r ≥ 0.57) than each of the quality features was to the talk-related quality dimension (r ≤ 0.27), the length of talk that was not qualified as rich (TA). However, for example, the association between talk-related and discursive dimensions seemed to be closer for the practice-based operationalizations (r = 0.27) than for the task-based operationalizations (r = 0.14).

To summarize, global statements about associations of quality dimensions cannot be provided as they fluctuated strongly with their operationalizations. Nevertheless, as the lowest correlations to the surface conceptualization of talk-related activation existed, we can at least infer that conceptual and discursive activation are in closer connection.

In contrast to the varying correlations in teachers' enacted activation, which indicated different patterns of small groups' engagement in the interaction, the correlations in the individual students' participation were much higher (Fig. 5). This was evident for the relations between talk-related and discursive (TP and DP-t; TP and DP-p), talk-related and conceptual (TP and CP-t; TP and CP-p), and discursive and conceptual participation (CP-t and DP-t; CP-p and DP-p). The lowest correlation (r = 0.67) was between the relative length of any individual talk not qualified in richness (TP) and the individual talk in conceptual practices (CP-p). Notwithstanding its strong connection, this shows that there was a difference when conceptualizing and operationalizing the quality and quantity of an individual students talk.

Hence, the methodological question arises in how far the individual students' discursive participation was already fully explained by the relative length of any individual talk not qualified in richness and conceptual participation, when operationalized in practice-based ways. As this question refers to the three quality features at the same time, it was pursued by multiple linear regression. The question of to what extent any individual talk time (TP) already explains the rich discursive participation is methodologically relevant because capturing all individual contributions (TP) does not necessitate extensive highly inferential coding as it does for the rich discourse practices (DP-p, CP-p).

Table 6 presents the results of the linear regression. It shows that individual students' participation in rich discourse practices (DP-p) was predicted by both students' individual conceptual participation (CP-p) and talk-related participation (TP) with a high explained variance R^2. The talk-related participation alone did not predict the discursive participation as well as the combined model. For capturing the quality of students' participation, it was thus important to conceptualize and operationalize all three quality dimensions.

5. Discussion

5.1. Summary

In this paper, we focused on a specific area of the deeper structure of instruction, namely interaction quality, by which we mean processes of the collective negotiating meanings of mathematical ideas (Bauersfeld, 1988; Walshaw & Anthony, 2008). Whereas there has been wide agreement that interaction quality matters for creating mathematically rich learning opportunities (Walshaw & Anthony, 2008), the methodological discourse on how to measure interaction quality and what to measure exactly is still ongoing, as the implicitness and heterogeneity of measurement approaches hinder a clear comparison of findings from different studies (Bostic et al.,

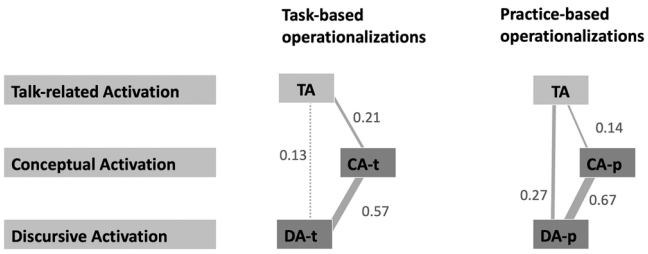


Fig. 4. Similar operationalizations—other conceptualization: Correlation between quality features for talk-related, conceptual, and discursive activation within task-based and practice-based operationalizations (non-significant correlations are depicted with dotted lines).

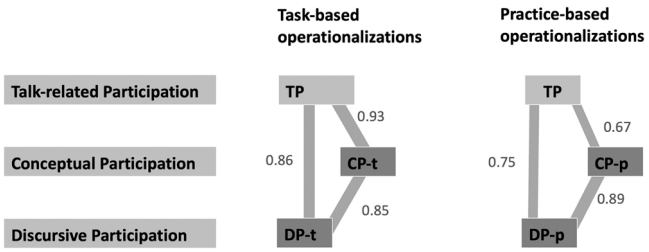


Fig. 5. Similar operationalization—other conceptualization: Correlation between quality features for talk-related, conceptual, and discursive participation within task-based and practice-based operationalizations (non-significant correlations are depicted in dotted lines).

Table 6
Multiple linear regression for predicting discursive participation (DP-p) by talk-related participation (TP) and conceptual participation (CP-p).

	Estimate (b)	Standard error	t-value	p
(Intercept)	0.002	0.002	1.233	0.21
CP-p	0.844	0.045	19.145	< 0.001
TP	0.169	0.023	7.428	< 0.001

$R^2 = 0.8415$, $F(2, 207) = 549.5$, $p < 0.001$

2021; Cai et al., 2020; Mu et al., 2022; Praetorius & Charalambous, 2018). As first examples have been given that conceptualizations and operationalizations can considerably influence the empirical results (Ing & Webb, 2012), our paper intends to contribute to this methodological discourse by offering constructs to describe and compare conceptualizations and operationalizations more systematically.

For making conceptualization decisions explicit, we systematized and studied different (often implicit) conceptualizations of interaction quality starting from three main quality domains identified in several case studies (see Lampert & Cobb, 2003): the pure space for student talk, mathematical (in particular conceptual) richness of the talk, and discursive richness of the talk. Furthermore, the focus of the conceptualization can be on teachers' intended activation or teachers' enacted activation (which has been shown not to be the same; Stigler et al., 1999) or individual students' participation, as their systematic investigation has been repeatedly identified as a research gap (Ing et al., 2015).

For contributing to a deeper methodological discourse on the impacts of operationalization decisions, quality features of existing video studies and coding protocols were analyzed and systematized. The literature review (in Sections 2.1, 2.2) revealed a large variety of heterogeneous quality features with task-based, move-based, or practice-based operationalizations and different ways for measurements (e.g., turn- or time-related rating or counting), often only roughly described in the reviewed papers. Based on the literature review, we narrowed down our analytic framework to six quality dimensions and several operationalizations for each, resulting in fourteen quality features that must be understood as only one (well-justified) example for different conceptualizations and operationalizations for fueling the methodological discourse.

Our empirical study contributes to the methodological discourse on both what to conceptualize and how to operationalize by disentangling different operationalizations in a video data set that was optimized for comparing interaction quality while keeping learning content, tasks, and representations constant. Although it might be theoretically plausible that by different conceptualizations and operationalizations different aspects of interaction quality are measured, this has sometimes been neglected in research design choices and transparency of reports (Mu et al., 2022). That the empirical evidence in this data set of 49 small-group classrooms supports the statement that methodological choices really matter is a finding of high methodological importance.

In all quality features, our empirical data show a large variety in teachers' enacted activations and individual students' participation and weak associations between them, which empirically underlines the request to treat them as different phenomena (Charalambous & Praetorius, 2018), and to capture individual students' participation using more than students' self-reports (as in TALIS; OECD, 2020).

Unlike the missing discourse might suggest, assessing a quality dimension with different operationalizations leads to different judgments (which strengthens the first evidence offered by Ing & Webb, 2012). The descriptive results on occurrence and deviations of the quality features in Table 4 already stress empirically that the different conceptualizations and operationalizations do not capture exactly the same phenomena. Theoretically, it is not astonishing that different conceptualizations can lead to different empirical results. However, it is remarkable that the choice of operationalizing quality features as task-, move- or practice-based leads to empirical results that differ strongly in their interpretation: In our data, the task-based features suggest a high interaction quality with students being engaged in conceptual activities in the task (CA-t is 77.4 %). But when assessing the enacted practices, the quality is judged much lower (CA-p is 23.5 %). This suggests that studying the correlations between the operationalized quality features might be of interest.

Within the same conceptualization in a quality dimension, we hypothesized and identified a closer relationship between the move-based and practice-based operationalizations than between task-based and practice-based operationalizations for conceptual activation (Fig. 2). This quantitative result resonates with case study findings (e.g., Walshaw & Anthony, 2008; Webb et al., 2008), since

working in mathematically rich, conceptual moves requires teachers to initiate and continuously support the class engagement in conceptual practices (Erath et al., 2018). Some studies have claimed a closer relationship between mathematically rich task demands and teacher moves when training teachers to implement those tasks in a specific way (Mata-Pereira & da Ponte, 2017). The existing association in our data, however, is only moderate when keeping the tasks constant across groups. The qualitative study by Franke et al. (2015) might offer an explanation: When teachers enact similarly rich mathematical move demands, this does not always coincide with similarly challenging discursive class engagement (see also in Cai et al., 2020; Walshaw & Anthony, 2008). Thus, when assessing the quality of conceptual activation, capturing teachers' move demands is not sufficient to capture all relevant differences in the interaction (see Quabeck & Erath, 2022, for a case study of two exemplary teachers whose enactment of conceptual practices considerably differed while enacting similar move-based quality). For the students' individual mathematical and discursive participation, we identified closer relationships between quality features with different operationalizations ($r = 0.67$) than for the teachers' enacted activation. Summing up, the low correlations between task-based operationalizations and the move- and practice-based operationalizations in teachers' enacted activation ($r \leq 0.2$; Section 4.2.2) indicate that in our data the measurement of interaction quality cannot be simplified to task-based operationalizations. In another publication, we will show that they vary also substantially in their power to predict students' learning gains (Prediger et al., in press).

When keeping the operationalization constant, correlations between different conceptualizations (talk-related, conceptual, and discursive quality dimensions) can be compared. When all operationalizations captured the same phenomena reliably, the connections between conceptualizations should be similar. Our data in Figs. 4 and 5 indicate that connections between conceptualizations are slightly different for task-based operationalizations than for practice-based operationalizations. To investigate in-depth the relationship between the quality dimensions (Decristan et al., 2020; Pauli & Reusser, 2015), we pursued a time-consuming practice-based operationalization by coding every student's utterance according to the richness of the discourse practices. As actively talking is a prerequisite for rich talking, it was expectable that the talk-related and discursive/conceptual participation were related, but the correlations in Figs. 4 and 5 quantify to what degree they are not the same. Both in the practice- and task-based operationalizations, the quality dimensions discursive and mathematical richness are in closer connection than their association to any talk in the talk-related quality dimensions. This was expected from the literature (e.g., Mercer et al., 1999; Sedova et al., 2019); however, our results provide quantified evidence that any talk differs strongly from mathematically and discursively rich talk. The multiple linear regression (Table 6) for predicting discursive participation (DP-p) from talk-related and conceptual participation (TP, CP-p) gives additional evidence that any talk (TP) has only a low explanation power for a student's individual participation in rich discourse practices. In contrast, CP-p explains a lot of the variance for DP-p that underlines the intertwinement of discursively and mathematically rich interaction (Erath et al., 2018).

5.2. Limitations

Of course, the empirical results presented and discussed must be interpreted with respect to the specificities of the data collected and analyzed. Our chosen conceptualizations and operationalizations, and hence the empirical analyses, depend on our broader understanding of mathematics learning, but we have tried to make this broader understanding transparent and comprehensible. Given the comparable conditions in the small-group teaching and nearly identical curriculum resources aiming at students' conceptual understanding, the empirical results detect more subtle differences between task-based, move-based, and practice-based operationalizations that might be less relevant in classes with stronger variation in declared learning goals, tasks, representations, and curriculum resources.

Furthermore, our small-group settings (with fewer students in total) increase the chances for individual students to participate in interactions as space for participation is only shared with (on average) four other students rather than more than 25 as in whole-class settings. In our study, we observed a relatively high number of quick switches in who was speaking, which coincided with contributions that were often brief in nature. This type of communication is different from what is typically expected in a whole-class setting, where it is more typical to have fewer contributions made by a smaller number of students (Erath et al., 2018; Sedova et al., 2019). Thus, in whole-class settings we would expect higher differences in individual talk-related, discursive, and conceptual participation when single students make longer contributions and others none. For the comparison of descriptive results (mean, SD), this is important: Students' talk-related participation evidently depends on the setting (7.7 % in time on task in this study versus on average 0.2 %, in Sedova et al., 2019). In addition, we would expect mathematically or linguistically less proficient students to have lower participation and engagement in classroom discussions in comparison to their more proficient peers (Lipowsky et al., 2007), which might have been addressed by the small-group teachers' more easily. Based on these specificities of the data, we can only speculate on how the small-group setting also influences the relation of quality dimensions, which should be studied in the future. Additionally, alternative conceptualizations and operationalizations should be studied in depth in order to increase the reliability of the findings.

The major limitation of our time measurements for moments of interaction above a defined quality bar (in conceptually or discursively rich tasks, moves, or practices) is that the underlying assumption (the more rich talk, the better for the learning) might be wrong. For the moment, this assumption is only justified by the qualitative analysis that provided insights into the powerful learning opportunities occurring in the moments of high-quality interaction (Webb et al., 2021; Quabeck & Erath, 2022). In the future, we intend to investigate the appropriateness of our assumption by analyzing how the different quality features influence students' learning gains, taking into account different student prerequisites (e.g., Bostic et al., 2021; Brophy, 2000; Decristan et al., 2020).

A further limitation is the choice of the selected frameworks for the literature review on possible operationalizations. Whereas most of the quality features identified stem from Bostic et al.'s (2021) review of valid coding protocols and therefore measurement of quality can be assured, the additional frameworks were added successively to fill "empty cells" in Table 2. Thus, we call for a more systematic

procedure in a follow-up study to insure that all the ways of operationalizing the nine conceptualizations of interaction quality are included.

Finally, a limitation of the current paper is that it did not yet provide evidence that the identified differences in capturing quality features also influence their ways of predicting students learning gains, because this would have overloaded the current paper. However, another publication will reveal that varying operationalizations and conceptualizations also have an impact on their predictive power for students' learning gains (Prediger et al., in press).

5.3. Implications for future research

In spite of its limitations, the current study can contribute to the ongoing methodological discourse about how to conceptualize and operationalize interaction quality in quantitative coding protocols with more care. The already raised concerns about conceptualizations and operationalizations (Cai et al., 2020; Praetorius & Charalambous, 2018) can be empirically substantiated by this study as has previously been done only by Ing and Webb (2012).

Of course, every quantitative coding protocol must necessarily simplify the complexities of interaction identified in qualitative studies (Lampert & Cobb, 2003). In the academic communication about coding protocols, however, the necessary and deliberate simplifications should be made more transparent, with the methodological choices being explicitly articulated and justified. Our constructs of quality domains (in our case space for talk, mathematical richness, and discursive richness; further quality domains should be added in the future), focus (teachers' intended activation, teachers' enacted activation, and student individual participation), and task-based, move-based, or practice-based operations offer a language to describe concisely the methodological choices taken in the coding protocols. They also do the same for measurement decisions (in our case, time-based measurements, but future study should investigate other measurements such as turn-based and frequency-based measurements). Empirical findings such as those of our and other studies (e.g., Ing & Webb, 2012) can help to justify each researcher's choices with more care.

For this reason, we also call for more studies that investigate alternative operationalizations in their coding protocols and account for different outcomes according to the operationalizations. This would strengthen the empirical foundations of methodological choices in the future and ensure higher comparability of the findings.

With these findings, we can provide an empirical base for the methodological concerns about the impacts of operationalizations raised by Praetorius and Charalambous (2018), Mu et al. (2022), and Cai et al. (2020). Future researchers should strive for deeper methodological discourses aiming at further exploring the impact of conceptualization and operationalization decisions for gaining comparable empirical results.

Declaration of Competing Interest

None.

Data availability

The authors do not have permission to share data.

Acknowledgment

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 (DFG-grants PR662/14-2 to S. Prediger, ER 880/3-3 to K. Erath).

References

Barwell, R. (2012). Discursive demands and equity in second language mathematics classroom. In B. A. Herbel-Eisenmann, J. Choppin, D. Wagner, & D. Pimm (Eds.), *Mathematics education library: Vol. 55. Equity in discourse for mathematics education: Theories, practices, and policies* (pp. 147–163). Springer.

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* (1st ed.,, pp. 27–46). NCTM and Lawrence Erlbaum Associates.

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

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

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

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

Brunner, E. (2018). Qualität von Mathematikunterricht: Eine Frage der Perspektive [Quality of Mathematics Instruction: A Question of Perspective]. *Journal für Mathematik-Didaktik, 39*(2), 257–284.

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

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

Charalambous, C. Y., & Litke, E. (2018). Studying instructional quality by using a content-specific lens: the case of the mathematical quality of instruction framework. *ZDM – Mathematics Education, 50*(3), 445–460. https://doi.org/10.1007/s11858-018-0913-9

Cohen, J. (1988). Statistical power analysis for the behavioral sciences. Erlbaum.

Decristan, J., Fauth, B., Heide, E. L., Locher, F. M., Troll, B., Kurucz, C., & Kunter, M. (2020). Individuelle Beteiligung am Unterrichtsgespräch in Grundschulklassen: Wer ist (nicht) beteiligt und welche Konsequenzen hat das für den Lernerfolg? [Students' differential participation in classroom discourse in primary schools: Who

participates (not), and what are the consequences for student learning?]. *Zeitschrift Für Pädagogische Psychologie, 34*(3–4), 171–186. https://doi.org/10.1024/1010-0652/a000251

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., Prediger, S., Quasthoff, U., & Heller, V. (2018). Discourse competence as important part of academic language proficiency in mathematics classrooms: the case of explaining to learn and learning to explain. *Educational Studies in Mathematics, 99*(2), 161–179. https://doi.org/10.1007/s10649-018-9830-7

Erath, K., Ingram, J., Moschkovich, J. N., & 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), 317–335. https://doi.org/10.1007/s11858-020-01213-2

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

Flick, L., Morell, P., Wainwright, C. (2004). Oregon Teacher Observation Protocol (OTOP). ⟨https://fg.ed.pacificu.edu/wainwright/Presentations/OTOP.Instrument.2005.Num.pdf⟩.

Franke, M. L., Turrou, A. C., Webb, N. M., Ing, M., Wong, J., Shin, N., & Fernandez, C. H. (2015). Student engagement with others' mathematical ideas. *The Elementary School Journal, 116*(1), 126–148. https://doi.org/10.1086/683174

Helmke, A. (2009). Unterrichtsqualität und Lehrerprofessionalität [Instructional Quality and Teacher Professionalism]. Kallmeyer.

Henningsen, M., & Stein, M. K. (1997). Mathematical tasks and student cognition: classroom-based factors that support and inhibit high-level mathematical thinking and reasoning. *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: A project of the National Council of Teachers of Mathematics* (pp. 371–404). Information Age Pub.

Hiebert, J., Gallimore, R., Garnier, H., Bogard Givvin, K., Hollingsworth, H., Jacobs, J., Miu-Ying Chui, A., Wearne, D., Smith, M.S., Kerstin, N., Manaster, A., Tseng, E., Etterbeek, W., Manaster, C., Gonzales, P., Stigler, J.W. (2003). Teaching mathematics in seven countries: Results from the TIMSS 1999 video study.

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

Hugener, I., Pauli, C., & Reusser, K. (Eds.). (2006). Dokumentation der Erhebungs- und Auswertungsinstrumente zur schweizerisch-deutschen Videostudie "Unterrichtsqualität, Lernverhalten und mathematisches Verständnis" [Documentation of the data collection and evaluation instruments for the Swiss-German video study "Teaching quality, learning behavior, and mathematical understanding"]. GFPF.

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

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

Jordan, A., Krauss, S., Löwen, K., Blum, W., Neubrand, M., Brunner, M., Kunter, M., & Baumert, J. (2008). Aufgaben im COACTIV-Projekt: Zeugnisse des kognitiven Aktivierungspotentials im deutschen Mathematikunterricht [Tasks in the COACTIV project: Report of the cognitive activation potential in German mathematics instruction]. *Journal für Mathematik-Didaktik, 29*(2), 83–107. https://doi.org/10.1007/BF03339055

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

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). National Council of Teachers of Mathematics.

Lipowsky, F., Rakoczy, K., Pauli, C., Reusser, K., & Klieme, E. (2007). Gleicher Unterricht – gleiche Chancen für alle? Die Verteilung von Schülerbeiträgen im Klassenunterricht [Same classroom – same chances for all? Distribution of student contributions]. *Unterrrichtswissenschaft, 35*(2), 125–147. https://doi.org/10.25656/01:5489

Marshall, J. C., Smart, J., & Horton, R. M. (2010). The design and validation of equip: An instrument to assess inquiry-based instruction. *International Journal of Science and Mathematics Education, 8*(2), 299–321. https://doi.org/10.1007/s10763-009-9174-y

Mata-Pereira, J., & da Ponte, J.-P. (2017). Enhancing students' mathematical reasoning in the classroom: teacher actions facilitating generalization and justification. *Educational Studies in Mathematics, 96*(2), 169–186. https://doi.org/10.1007/s10649-017-9773-4

Mercer, N., Wegerif, R., & Dawes, L. (1999). Children's talk and the development of reasoning in the classroom. *British Educational Research Journal, 25*(1), 95–111. https://doi.org/10.1080/0141192990250107

Moschkovich, J. (1999). Supporting the Participation of English Language Learners in Mathematical Discussions. *For the Learning of Mathematics, 19*(1), 11–19.

Moschkovich, J. N. (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

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

Neubrand, M., Jordan, A., Krauss, S., Blum, W., & Löwen, K. (2013). Task analysis in COACTIV: Examining the potential for cognitive activation in German mathematics classrooms. In M. Kunter, J. Baumert, W. Blum, U. Klusmann, S. Krauss, & M. Neubrand (Eds.), *Cognitive Activation in the Mathematics Classroom and Professional Competence of Teachers: Results from the COACTIV Project* (pp. 125–144). Springer.

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

OECD. (2020). *Global Teaching InSights: A Video Study of Teaching*. OECD Publishing,. https://doi.org/10.1787/20d6f36b-en

Pauli, C., & Lipowsky, F. (2007). Mitmachen oder zuhören? Mündliche Schülerinnen- und Schülerbeteiligung im Mathematikunterricht. *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: a video-based study on mathematics teaching in German and Swiss classrooms. In L. B. Resnick, C. S. C. Asterhan, & S. N. Clarke (Eds.), *Socializing Intelligence Through Academic Talk and Dialogue* (pp. 181–193). AERA.

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

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., Erath, K., Quabeck, K., & Stahnke, R. (in press). Effects of interaction qualities beyond task quality: Disentangling instructional support and cognitive demands International Journal of Science and Mathematics Education.

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. (2022). Measuring interaction quality: how much detail is necessary? Results from a quantitative video study on the conceptual dimension. In J. Hodgen, E. Geraniou, G. Bolondi, & F. Ferretti (Eds.), *Proceedings of Twelfth Congress of the European Society for Research in Mathematics Education (CERME 12) (8 pages)*. University of Bolzano/ERME. ⟨https://hal.archives-ouvertes.fr/hal-03746019⟩.

Sawada, D., Piburn, M. D., Judson, E., Turley, J., Falconer, K., Benford, R., & Bloom, I. (2002). Measuring reform practices in science and mathematics classrooms: the reformed teaching observation protocol. *School Science and Mathematics, 102*(6), 245–253. https://doi.org/10.1111/j.1949-8594.2002.tb17883.x

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. (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. (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

Schwarz, C. V., Braaten, M., Haverly, C., & los Santos, E. X. de (2021). Using sense-making moments to understand how elementary teachers' interactions expand, maintain, or shut down sense-making in science. *Cognition and Instruction, 39*(2), 113–148. https://doi.org/10.1080/07370008.2020.1763349

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

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

Stigler, J.W., Gonzales, P., Kawanaka, T., Knoll, S., Serrano, A. (1999). The TIMSS Videotape Classroom Study: Methods and findings from an exploratory research project on eighth-grade mathematics instruction in Germany, Japan, and the United States. National Center for Education Statistics.

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

Webb, N. M., Franke, M. L., Johnson, N. C., Ing, M., & Zimmerman, J. (2021). Learning through explaining and engaging with others' mathematical ideas. *Mathematical Thinking and Learning*, 1–27. https://doi.org/10.1080/10986065.2021.1990744

Webb, N. M., Franke, M. L., Ing, M., Chan, A., De, T., Freund, D., & Battey, D. (2008). The role of teacher instructional practices in student collaboration. *Contemporary Educational Psychology, 33*(3), 360–381. https://doi.org/10.1016/j.cedpsych.2008.05.003

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