Learning and Instruction 86 (2023) 101777

Contents lists available at ScienceDirect

Learning and Instruction

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

The role of interactive dialogue in students' learning of mathematical reasoning: A quantitative multi-method analysis of feedback episodes

Robbert Smit [a,*], Kurt Hess [b], Alexandra Taras [a], Patricia Bachmann [a], Heidi Dober [b]

[a] St. Gallen University of Teacher Education, Notkerstr. 27, 9000, St.Gallen, Switzerland
[b] University of Teacher Education Zug, Zugerbergstr. 3, 6300, Zug, Switzerland

ARTICLE INFO

Keywords:
Formative feedback
Mathematical reasoning
Primary school
Interactive dialogue

ABSTRACT

Interactive dialogue within feedback episodes is essential for developing primary school students' mathematical reasoning competence. Our goal was to better understand the nature of the associations between observed dialogue, teachers' formative feedback, and students' mathematical reasoning. We applied a two-step approach, first constructing a video-analysis instrument for assessing the quality of interactive dialogues and then combining the interaction data with student and teacher questionnaire data from 804 students in 44 fifth and sixth grade primary school classes. The quality of the observed dialogues predicted class differences in students' self-efficacy for explaining but not in their reasoning competence, which was predicted by perceived formative feedback.

1. Introduction

Dialogue is essential for developing school students' mathematical reasoning skills (Howe, Hennessy, Mercer, Vrikki, & Wheatley, 2019). Participating in mathematical discussions that involve conjecturing, explaining, justifying, and evaluating solutions is a typical learning activity in the mathematics community, and it is not reserved only for reasoning (Moschkovich, 2007). Mathematical reasoning is a special kind of discussion or dialogue in which the goal is to determine the truth of mathematical statements. Learning as co-construction requires the teacher and student(s) to be partners in exchanging arguments in dialogic interactions, although the teacher is, of course, the expert (Adie, Kleij, & Cumming, 2018). This socio-constructive perspective of learning is related to Vygotsky (1978), who viewed knowledge and understanding as social constructions created through interactions with others.

Verbal feedback from teachers is a key element of supporting students in making claims or finding arguments as part of such dialogue (Fyfe & Brown, 2018). This feedback support for reasoning can occur during discussions among an entire class or individual dialogue between a student and teacher. In this study, we investigated the latter, that is, feedback dialogue as interactions with individuals or small groups of students (mostly pairs) during seatwork. Given that 'feedback involves a reciprocal and dialogic process of [the] co-construction of meaning' (Hattie & Gan, 2011, p. 251), it does not merely involve the teacher telling the students which steps to take or correcting their mistakes; feedback includes the teacher encouraging the students to think further about what has been discussed. Hence, it is not the classical feedback sequence of initiation-response-feedback but rather the process of the teacher listening to the students to understand their thinking with respect to a predetermined learning goal (Davis, 1997; Lim, Lee, Tyson, Kim, & Kim, 2020). The teacher's role is to provide a space for students to become aware of mathematical relationships, to ask questions, to develop their argumentation skills, and to articulate their reasoning. For students to learn reasoning well, the teacher must provide students with feedback on the accuracy and appropriateness of the language they use to explain and argue mathematically (Bragg, Herbert, Loong, Vale, & Widjaja, 2016). This activity implies the need for a careful balance between the pupil's independent progress and the teacher's feedback and scaffolding (Howe & Abedin, 2013), and providing learners with opportunities to submit and receive feedback on school work that has not been assessed is part of formative assessment (Winstone, Nash, Parker, & Rowntree, 2017). In the context of formative feedback, an interactive dialogue can be viewed as an assessment via conversation, in which the teacher continually acquires information about the student's level of linguistic expression and level of understanding of the topic

* Corresponding author.
E-mail addresses: robbert.smit@phsg.ch (R. Smit), kurt.hess@phzg.ch (K. Hess), alexandra.taras@phsg.ch (A. Taras), patricia.bachmann@phsg.ch (P. Bachmann), heidi.dober@phzg.ch (H. Dober).

https://doi.org/10.1016/j.learninstruc.2023.101777
Received 26 December 2022; Received in revised form 22 March 2023; Accepted 27 March 2023
Available online 10 April 2023

(Ruiz-Primo, 2011). In the dialogic classroom framework, our study is situated in the interactive one-to-one setting (Alexander, 2018).

Teachers who view teaching mathematics primarily as acquiring mathematical concepts and operations often struggle to help students develop competences in problem-solving or logical thinking (Schoenfeld, 2015). The challenge is making sense of students' approaches so that they can learn from both wrong and right ideas. Regarding mathematical reasoning, straightforward approaches do not always exist, leaving some students uncertain about their answers. These students often have low mathematics self-efficacy (Tait-McCutcheon, 2008; Tanner & Jones, 2003), and classroom observations show that comforting feedback from the teacher, in combination with praise, can help them persist in their learning tasks (Eriksson, Björklund Boistrup, & Thornberg, 2017). Efficacy beliefs have also been reported to mediate feedback effects on mathematics achievement (Rakoczy et al., 2019). In relation to problem-solving and self-efficacy beliefs, students should learn to apply self-regulation strategies and to assume greater responsibility for their own learning (Butler & Winne, 1995; Van der Schaaf, Baartman, Prins, Oosterbaan, & Schaap, 2013).

According to Howe et al. (2019), classroom dialogue has been examined mostly among students working independently of teachers in small-group interactions, and little is known about the dialogue between teachers and individual students or small groups. A recent video study by Stovner and Klette (2022), which investigated classroom feedback practices in a Norwegian secondary school, showed that almost no feedback was given related to mathematical practices, such as problem-solving or justifying mathematical claims. Lim et al. (2020) also concluded that more classroom research is needed to examine which teacher feedback practices foster mathematics discussions and to reveal how students perceive mathematical practices while engaging in teacher-student dialogues and co-constructing mathematical understanding.

Interactions in dialogue, formative feedback, and learning have been investigated mostly in higher education settings and less often in primary school classrooms (Meusen-Beekman, Joosten-ten Brinke, & Boshuizen, 2016; Stovner & Klette, 2022; Winstone et al., 2017). Regarding the secondary classroom, Ruiz-Primo and Furtak (2007) suggested that effective informal formative assessment practices might be associated with student learning in scientific inquiry classrooms. However, they studied only interactions between the teacher and the class. Although in a recent study using student data we showed that formative feedback predicted mathematical reasoning mediated by students' self-efficacy beliefs (Smit, Dober, Hess, Bachmann, & Birri, 2022), we did not include teacher data and external observations. Therefore, the goal of the present study was to examine the associations between observed interactive dialogues between teachers and students, teachers' formative feedback practices, and students' mathematical reasoning competence in upper primary school classes (grade 5/6, age 11/12). We applied a two-step approach to gathering empirical evidence: first, we constructed a video-analysis instrument to assess the quality of the interactive dialogues observed in the classroom, and second, we combined the feedback interaction data with questionnaire data collected from students and teachers.

1.1. Interactive dialogues

In the educational research literature, the term 'classroom dialogue' has different conceptualisations and terminology, such as accountable talk, dialogic inquiry, exploratory talk, and dialogic teaching (Alexander, 2018; Hennessy et al., 2016). In this study, interactive dialogue – a specific form of classroom talk – is associated with individual feedback conversations between a teacher and pupil working on a task, not classroom conversations among the entire class. In the framework of dialogic teaching, these interactive settings are one-to-one settings. Although dialogue also includes the notion of reaching a shared understanding, which is important for mathematical reasoning, classroom talk can be basic conversations as well. In this paper, we report research findings on classroom dialogue in general because published research on interactive one-to-one dialogues related to mathematical reasoning in our setting – between teacher and student during seatwork – is scarce.

Howe and Abedin (2013) note that classroom dialogue is rooted in philosophical and pedagogical treatises, and they claim that the exchange of ideas is highly beneficial for learning. The term 'interactive' underscores the importance of the teacher's use of questions that elicit responses from the students (Chi, Siler, Jeong, Yamauchi, & Hausmann, 2001). In monologic dialogue, the teacher explains information in detail without responding to students' ideas, and interaction is limited (Drageset, 2014). The typical skills teachers need for one-to-one dialogue include giving feedback, scaffolding, providing explanations, knowing when to refrain from correcting errors, and allowing students to infer that an error has been made (Chi et al., 2001). Chi et al. (2001) indicate that although teachers generally possess domain knowledge, they do not necessarily have the skills needed for interactive dialogue. In exploring the role of certain teaching moves to enhance learning within teacher-student dialogue, they discovered in their study that for students, responding to (the teacher's) scaffolding and reflection contributed to the effectiveness of dialogic interactions. However, for teachers, neither explanations nor comprehension-gauging questions boosted learning, and scaffolding correlated only with shallow learning. Chi et al. (2001) showed that students learned just as well when they did not hear any teacher explanations or receive one-way feedback.

In the context of classroom dialogues, Lim et al. (2020) and Drageset (2014, 2015) investigated teachers' discourse practices related to mathematical reasoning. Drageset (2015) identified teacher-discourse turns, in which the teacher questioned primary students to acquire further insights into their thinking, as opposed to turns, in which the teacher posed questions to check students' understanding. The study by Lim et al. (2020), which examined secondary school students, found that the mathematics discussions the students found most fruitful were those in which the teachers tended to ask questions and then listen and respond to the students' ideas or reasoning. The challenge for the teacher in such discussions is to grasp the students' reasoning in sufficient depth to provide appropriate feedback (Howe & Abedin, 2013).

1.2. Interactive dialogue and its relationship to formative feedback, self-regulation, and efficacy beliefs

Interactive dialogues allow teachers to explain their feedback, provide examples, and explore ways to improve learning. Hattie and Gan (2011, p. 251) refer to feedback as a 'social negotiation through the development of cognitive and evaluative skills in the development of understanding'. Here, we find a mutual understanding of the necessary social interaction between teacher and student for effectively guiding learners, as occurs in interactive dialogues. During students' individual work on tasks, interactive dialogues provide opportunities for teachers to conduct formative assessments of their understanding. In the Wiliam and Thompson (2007, p. 63) model, 'feedback that moves learners forward' is an important component of formative assessment. The aim of formative assessment is to gather information generated by learning activities and artefacts to adjust instruction and promote learning (Black & Wiliam, 2009). Feedback becomes formative when 'it is used to [close] the gap between the actual level and the reference level of a system parameter' (Ramaprasad, 1983, p. 4). In school, this parameter is the learning objective in the lesson. In addition, 'for feedback to be formative it should be an episode of learning for both the student and the teacher', and teachers 'should concentrate on using a feedback episode to learn about students' thinking and understanding' (Ruiz-Primo & Brookhart, 2018, p. 51). The use of the term 'episode' underscores the characteristic of feedback, which is not a one-sided act by the teacher but might consist of a longer talk in which the teacher starts with questions, listens to the student's thoughts, interprets their implications, and then provides feedback that moves the student forward.

As reported in a study on secondary students by Van der Schaaf et al. (2013), feedback in the form of dialogue is more effective than feedback in writing only. According to Van der Schaaf et al. (2013), to benefit most from formative feedback, students must make connections between feedback they receive and their future actions on the task. Thus, reflective thinking and constructive responses from the students are required (Chi et al., 2001; Wu & Schunn, 2021). As part of their reflection, learners use information gained from the completed task to improve their understanding, explanations, and discussions in their next task (Shilo & Kramarski, 2019). Because reflective or metacognitive thinking involves active selection and uses thinking activities to perform complex tasks, it is conceptualised as a component of students' self-regulation (Boekaerts, 1999). In models of self-regulation interactions are described between monitoring processes and controlling actions. These interactions are often concretised as problem-solving strategies for specific tasks (Butler & Winne, 1995).

Self-efficacy refers to students' beliefs that they are capable of performing the behaviours necessary to achieve certain goals (Bandura, 1977). It has been theoretically posited that students' mathematics self-efficacy can affect their mathematics achievement by influencing certain behavioural and psychological processes (Bandura, 1997). Students with low efficacy beliefs can benefit from cautious interactive dialogue in which teachers try to understand the students' reasoning without immediately calling it wrong (Hattie & Clarke, 2019). Peers also provide vicarious experiences and are important resources for feedback, and both of these characteristics are important sources for self-efficacy beliefs (Bandura, 1977).

However, feedback does not work equally well for everyone. Fyfe and Brown (2018) reviewed eight experimental studies on the effects of feedback on children's reasoning about math equivalence. Although the students (ages 6–11) with little prior mathematics knowledge seemed to profit from feedback during mathematical reasoning, those with some prior mathematics knowledge did not benefit from feedback during problem-solving, and they experienced significant negative effects, leading to lower scores on a test than students who had not received any feedback (Fyfe & Brown, 2018).

1.3. Mathematical reasoning and its challenges in the classroom

What 'mathematical reasoning' means has not yet been definitively clarified, even in the research community (Jeannotte & Kieran, 2017). In addition to 'reasoning', the term 'argumentation' is also often used without any real distinction (Whitenack & Yackel, 2002), and the two terms do not have well-defined boundaries (Hanna, 2014). According to Viholainen (2011), reasoning can be defined as a process in which arguments are exchanged with the aim of reaching a convincing conclusion. Similarly, Lithner (2000) describes reasoning as a four-stage process consisting of a problematic situation, the choice of a strategy, its application, and a conclusion. Reasoning originates from students going through these stages, linking them, and deriving a final rationale. In examining concrete activities, it could be argued that argumentation and reasoning refer to the same mathematical process. In most research, the term 'reasoning' is used, but in the Swiss Mathematics Standards (Swiss Conference of Cantonal Ministers of Education (EDK), 2011), which form the basis for our study, 'reasoning' and 'argumentation' are used synonymously. The standard for students at the end of their primary school education (grade six) states: 'Argumentation and reasoning require the ability to verify statements and justify or falsify results using data or arguments' (Swiss Conference of Cantonal Ministers of Education (EDK), 2011, p. 40). Thus, a student should be able to justify whether a mathematical claim is either generally true or true only under certain conditions. The following example illustrates what is expected from a student at the end of primary school:

> 0.8, 0.88, 0.888, 0.8888, …
>
> In this sequence, is a number always larger than the preceding number? Give reasons for your answer!

According to Stylianides (2008), formal proofs are also part of reasoning, but they are generally not listed in primary and lower-secondary school curricula and are only introduced in upper-secondary school lessons. However, building argumentation skills as part of preliminary stages of formal reasoning can begin as early as primary school (Blum & Kirsch, 1991; Semadeni, 1984). Such teaching involves explaining the procedure, assumptions, or results regarding formulating assertions, predictions, and generalisations (Bezold, 2009). Specifically for algebraic reasoning, an example in the lower grades could involve exploring patterns and describing relationships (Lüken, Peter-Koop, & Kollhoff, 2014).

A special feature of reasoning tasks is that they often take the form of word problems (Cummins, Kintsch, Reusser, & Weimer, 1988), meaning that the task is embedded in a short text with one or more questions (Verschaffel, Schukajlow, Star, & Van Dooren, 2020). So it follows that acquiring reasoning also requires and can promote language skills in students (Bragg et al., 2016). Such tasks often contain illustrations, such as tables, with further information.

Teaching reasoning also includes incorporating pedagogical considerations to facilitate communication among children. These factors include more frequent group work, writing, project work, and extended forms of assessment. In other words, teachers should have the understanding that in order to use reasoning tasks, it is of great importance to create a classroom environment in which the focus is on classroom discourse and one-to-one dialogues between teachers and students. In addition, discursive exchange in such an environment would allow teachers to gain insight into the students' thought processes, making it crucial for understanding learning processes (Brodie, 2010; Ginsburg, 2009; Sfard, 2001).

1.4. The present study: research questions and hypotheses

In this exploratory research, we first aimed to use video analysis to measure the quality of teachers' interaction dialogue within feedback episodes for mathematical reasoning. The objective of this first step was to clarify which aspects of interactive dialogue in giving feedback tended to be easy or frequent and which were difficult or infrequent. Second, we used questionnaires to determine whether the teachers' interactive dialogues were consistent with their students' competence in mathematical argumentation. Drawing on insights and results from a theoretical model by Panadero and Jonsson (2013), we expected to find indirect effects of interactive dialogue between teachers and students – as shown in Fig. 1 – via the teachers' formative feedback within this dialogue and the students' self-regulation and self-efficacy beliefs. Panadero and Jonsson developed their theoretical model based on a literature review on the effects of formative rubrics for student achievement. Rubrics aid the feedback process by providing criteria or transparency of the goals, resulting in various beneficial factors related to student achievement, such as better self-efficacy or self-regulation. In two previous studies using student questionnaire data, we found support for the relationship between feedback and student beliefs (Smit et al., 2017, 2022).

In the study presented here, we analysed the effects at the individual/student and clustered/class levels. Hence, we were able to comment on differences in reasoning competence between and within classes. In addition, we combined different views of teaching: teacher and student perceptions were supplemented by the objective views of experts, which provided a more complete picture of what was happening in the classroom.

Our model and expectations led to the following hypotheses:

Hypothesis 1. We expected the quality of the teachers' interactive dialogues to have an indirect effect on the students' competence in mathematical reasoning via the students' perceived self-efficacy in

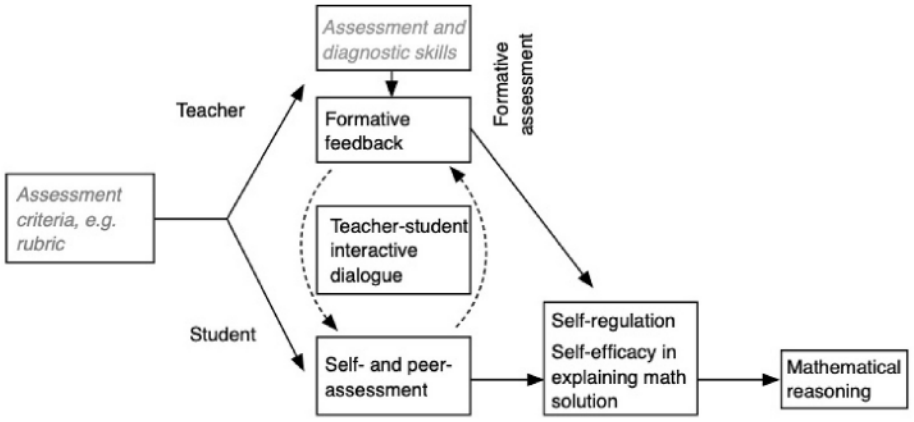

Fig. 1. The research model was derived from the project 'Learning with Rubrics' (Smit et al., 2017). The variables in italics/grey are not part of this study.

explaining skills.

Hypothesis 2. We expected the quality of the teachers' interaction dialogues to have an indirect effect on the students' competence in mathematical reasoning via the students' perceived self-regulation skills for word problems.

Hypothesis 3. We expected the quality of the teachers' interaction dialogues to be related to the students' perceptions of the teachers' formative feedback practices and the teachers' use of peer feedback.

Hypothesis 4. We expected the teachers' formative feedback to have an indirect effect on the students' competence in mathematical reasoning via the students' perceived self-efficacy in explaining and self-regulation skills.

2. Methods

2.1. Design

This observational study is part of a project titled 'Feedback for Mathematical Reasoning', which was conducted by two teacher-education universities in Switzerland (St. Gallen and Zug). The project began in 2018 and ended in 2022. Approximately half of the data were derived from a prior, linked project titled 'Learning with Rubrics' (Smit et al., 2017), which was based on an identical research design (Table 1). In the previous project (2014–2016), only some of the participating teachers allowed previously not included classroom videotaping.

We used a survey with a longitudinal pretest-posttest design and video data retrieved from lessons in which students worked on reasoning tasks. In addition, we measured the students' competence in mathematical reasoning at the beginning of the implementation phase (T1). During the same week, the students completed a questionnaire on their attitudes and aspects related to classroom learning, such as formative assessment practices. Towards the end of the implementation phase (lesson 8 or 9), a lesson sequence for students working individually or in small groups on mathematical reasoning tasks was recorded in each class (around 30 min). The teacher was requested to provide individual support for students based on their level of learning. Although we scheduled the videotaped lessons using the same series of reasoning tasks, some recordings contained different reasoning tasks than those planned. Some teachers were behind schedule, and it was not always possible for the research team to visit classes at the optimal time because of time constraints.

At the end of the implementation phase, we measured the students' competence, attitudes, and perceptions of classroom learning again (T2). In this study, we focused on the measurements taken at the end of the project, as it was only then that all the students could have acquired the same understanding of mathematical reasoning, as provided by the tasks and the teachers' explanations.

Table 1
Research design.

T1 Pretest & Questionnaire August 2019	Training of mathematical reasoning in the classroom	T2 Posttest & Questionnaire November 2019	Evaluation November 2019
Teacher questionnaires • Formative feedback • Self- and peer assessments	Teachers participate in a workshop on mathematical reasoning. Training phase in the classroom.	Teacher questionnaires • Identical to T1	Evaluation of the implementation of mathematical reasoning in the classroom
Student questionnaires and test of mathematical reasoning • Formative feedback • Self-regulation word problem • Self-efficacy explaining solutions • Reasoning test	Video recording of one lesson during lesson 8 or 9 – interactive dialogue	Student questionnaires and test of mathematical reasoning • Identical to T1	

Note: T1 = Beginning of the implementation phase; T2 = End of the implementation phase.

2.2. Participants

Teachers and their classes were recruited with the help of an advertisement in a local teachers' journal, in addition to personal requests for participation. After the first round of our project in 2015, which included 45 participants, we repeated the recruitment process in 2019, enrolling another 28 teachers. However, only 44 teachers were willing to participate in the classroom videotaping. Thus, our sample for this study consisted of 44 full-time teachers and their classes from two regions in central and eastern Switzerland. Of these individuals, 27 were women and 17 were men, their mean age was 39 years, and their mean length of service was 13 years. Twenty-one teachers managed a fifth-grade class, 13 were responsible for a sixth-grade class, and the remaining 10 teachers taught multi-grade classes. Two of the remaining 10 teachers also involved their fourth-grade students in the project. Thus, we obtained 44 class datasets at two points in time, consisting of 42 fourth graders, 454 fifth graders, and 308 sixth graders (804 students total). Approximately 51% of the students were boys and 49% were girls, and almost all the students were white and identified as Swiss. The

official school language in the region of the study was German, but German was not the mother tongue for 26% of the students. Data collection followed ethical standards of the swissethics committee and was in accordance with the declaration of Helsinki. Informed consent was obtained from the participants in the study.

2.3. Procedure

The same team members conducted separate workshops at each of the two teacher-education universities. During these workshops, theoretical knowledge related to the content and teaching of mathematical reasoning was presented. The teachers' previous personal experiences working on reasoning tasks enabled a fruitful discussion of the implications for teaching. The implementation phase, which followed the workshop, consisted of nine weeks of training for the students, and one lesson on practicing mathematical reasoning was delivered each week. All teachers were required to follow a strict script from a detailed lesson plan (Table 2). These lesson plans had a socio-constructivist orientation, meaning that the learners constructed new knowledge through engagement in activities and mathematical reasoning within a community (Ball & Bass, 2000). A collaborative group or peer work activity (e.g. placemat activities) to enhance student dialogue was part of almost every lesson (Knudsen, Lara-Meloy, Stevens Stallworth, & Wise Rutstein, 2014; Mercer & Sams, 2006).

During the first part of the script, the teachers were asked to discuss the quality of different examples of reasoning with the class to clarify the targets. During lesson implementation, the students engaged in peer- and self-assessments of their work, especially towards the end of the lesson series, when it was important to practise what they had learned on an individual level. During the final weeks, the teachers were expected to give each student feedback on their level of competence in mathematical argumentation. The teachers received a set of exercises for mathematical argumentation, including possible solutions. These exercises covered numerical sequences, quantities and units, the decimal system, basic operations, proportions, and estimations. Some of these tasks were purely mathematical, whereas others were contextualised and related to reality.

2.4. Instruments

2.4.1. Instrument for rating the interactive dialogue within feedback episodes

For rating the video material, we developed a coding instrument. Our coding instrument, which consisted of 14 categories, was designed to estimate the quality of the interactive dialogues of these feedback episodes (see Table 3). The first category – an interaction with several turns – served as the criterion for the inclusion or exclusion of the previously identified feedback episode (Chi et al., 2001). These sequences of interactive dialogue with different durations were either a dialogue with one student and a teacher or with a small group (mostly pairs) of students and a teacher. The instrument was used to collect observational data on the teachers' and students' behaviour during interactive dialogue. Although teachers and students were both needed for sustaining the dialogue, they had different roles, such as the teachers providing feedback and the students posing questions or explaining their approaches for mathematical reasoning.

To develop our instrument, we used four papers relevant to our focus on interactive dialogue as part of teachers' feedback on students' mathematical reasoning. We also analysed other articles, such as Alexander et al. (2017); Ruiz-Primo (2007); Webb et al. (2017), to identify significant features of interactive dialogue or feedback, but we found the initial four articles sufficient for our instrument. The aim was not to select a complete set of items but enough that could reliably measure teacher competence. Three of the articles (Adie et al., 2018; Chi et al., 2001; Howe & Abedin, 2013) pertained to the effectiveness of interaction or classroom dialogue, including feedback and explanations, and one article focused on mathematical reasoning (Bragg et al., 2016). Chi et al. - although a slightly older study - was chosen because they examined one-to-one scaffolding situations.

A useful article by Hennessy, Howe, Mercer, and Vrikki (2020) was published after this study was conducted. Following their three coding levels for classroom dialogue (micro-, meso-, and macro-levels), we rated the longer episodes at the macro-level. However, when specific situations (as defined by our manual) occurred during dialogic interactions, our ratings included the lower levels.

2.4.2. Questionnaires

We used two questionnaires to collect data on 1. teachers' and 2. students' attitudes and their perceptions of the class. Our research model guided us through our choice of assessment scales (Fig. 1). We operationalised the constructs carefully, according to theory, and then adapted them to the specific needs and context of the study. Hence, construct validity was ensured by selecting the best measures available. When such measures did not exist, we constructed new items appropriate for the construct and targeted sample, using theoretical knowledge and teaching experience. The students' understanding of the items was piloted using a small sample.

Our model for the choice of the three assessment scales for the teachers' questionnaire was based on the theoretical work of Black and Wiliam (2009, p. 8). We constructed scales for the teacher questionnaires, which were based on existing items (Brown, Harris, & Harnett, 2012; Smit, 2009) and items adapted from the research literature (Hargreaves, McCallum, & Gipps, 2000; Hattie & Timperley, 2007). The background variables included the teachers' gender and teaching experience.

The three scales and items from the student questionnaire used in this study are presented in detail in the Appendix. The student questionnaire was used to collect information on nationality, gender, and family background (e.g. topics of books at home; see Verhoeven and van Elsäcker (2016)).

For each scale, the score was calculated as the average of the grand mean values of the variables.

Diagnostic skills (teacher questionnaire). Teachers' diagnostic skills serve the function of formative assessment. According to Black and Wiliam (2009), a diagnosis must precede the formulation of any feedback and should inform the teacher about the student's thinking and motivation. This process includes diagnoses of misconceptions, misinterpretations, confusion, and difficulty demonstrating a skill or a concept that has not been developed. A sample item is 'It is easy for me to identify pupils' learning difficulties in mathematics.' Answers were reported using a Likert scale ranging from 1 (do not agree at all) to 6 (fully agree). The internal consistency of the scale, as measured by Cronbach's α, was 0.79.

Formative feedback (teacher questionnaire). Formative feedback is the information provided to students about the quality of their work to help them make the necessary adjustments to improve their performance. The feedback should be useful for the student (Ruiz-Primo & Brookhart, 2018, p. 128). Hattie and Timperley (2007) identified four levels: feedback on (1) the self, (2) the task, (3) the process, and (4) self-regulation levels. We chose three of the four most beneficial levels. A sample item is 'During the task I discuss with the pupils whether their learning is going in the right direction.' Our scale consisted of eight items related to these three levels of feedback. Answers were reported using a Likert scale ranging from 1 (never) to 6 (always). The internal consistency of the scale, as measured by Cronbach's α, was 0.72.

Self- and peer-assessments (teacher questionnaire). Self-assessments help students monitor where they are in relation to their learning goals and plan where they will go next (Ruiz-Primo & Brookhart, 2018). In this context, self-assessment is a functional part of formative assessment. Peer-feedback differs from self-assessment because the source is external. Nevertheless, providing feedback to other students is similar to self-assessment, and it is an instructional

Table 2
Lesson plan (overview).

Lesson 1	Lesson 2	Lesson 3	Lesson 4	Lesson 5	Lesson 6	Lesson 7	Lesson 8	Lesson 9	Lesson 10	Lesson 11
Pretest	Introduce mathematical reasoning.	Note and evaluate good reasoning.	Solve reasoning tasks.	Solve argumentation tasks, teacher feedback.	Solve argumentation tasks, teacher feedback.	Present re-entry lesson after autumn break.	Repetition: Optimise the argumentation of the solution.	Practise: Solve and assess tasks, feedback from teacher.	Practice/ conclusion: Solve and assess tasks, feedback from teacher.	Posttest
TASKS FOR STUDENTS (OR TEACHER)	Work on initial task in groups (placemat method), describe and justify solutions.	In class: Sort and assess sample solutions from the last lesson according to quality.	Assess anonymous solution of a task and define optimisation suggestions in groups (placemat method).	In class: Work out strategies for working on reasoning tasks based on a sample solution.	Solve tasks in groups, self-assess and give feedback to other groups (placemat method). Check and implement optimisation suggestions.	Recapitulate 'good mathematical reasoning'.	Provide input (teachers), review findings from the previous lessons.	Present a good mathematical reasoning task of last lesson (teachers).	Evaluate the tasks by others completed in the last lesson (group work).	
	In class: Collect elements of a convincing argumentation and solve another task.	Work on further reasoning tasks with self-assessment and peer feedback.	In class: List central optimisation suggestions.	Solve tasks individually, self-assessment, peer feedback, implement optimisation suggestions (learning tempo duet method), note personal points to remember.	Individual teacher feedback for the second half of the class.	Formative assessment, including self-assessment.	Solve tasks in groups and assess solutions of other groups (placemat method), note suggestions for improvement.	Solve further tasks individually and self-assess, feedback in partner work (learning tempo duet method).	Give feedback to others on particularly successful tasks.	
	Input by teacher: Clarify elements and quality features of reasoning.	Implement optimisation suggestions (learning tempo duet method).	Solve and assess tasks individually (self-assessment), implement optimisation suggestions.	Individual teacher feedback for approx. half of the class.		Complete tasks from previous lessons.		Individual teacher feedback for approx. half of the class.	Solve further tasks individually and self-assess them, give feedback in partner work (learning tempo duet method).	

Table 3
Categories of teachers' interactive dialogue within feedback episodes.

Number	Category	Statements from sources	Source
0	Interactive dialogue has several turns	However, in tutoring, two additional steps are taken. In the fourth step, the tutor proceeds with a scaffolding episode, one that can require 5–10 turns.	Chi, M.T.H., Siler, S.A., Jeong, H., Yamauchi, T., & Hausmann, R.G. (2001), p. 473.
1	Teacher provides only wanted feedback	Dialogic interactions require a shift in traditional teacher–student relationships, moving away from teachers as providers of feedback and telling students the next steps to take, to empowering students in their learning as partners in co-constructing meaning. In contrast, when teachers act as 'advice dispensers' and 'solution providers', feedback 'serves as an external evaluation and can be described as a "gift" that may be neither wanted nor acted upon. This often-uninvited form of feedback may not necessarily be the learner's focus'.	Adie, L., Kleij, F., & Cumming, J. (2018), p. 706.
2	Prompts by teacher are open-ended and content-free	These [teacher] prompts were designed with several goals and constraints in mind. First, they were open-ended, which seemed to be more conducive to inviting responses from the students, such as 'What's going on here?', 'Anything else to say about it?', 'Could you explain or put this in your own words?', and 'What do you think?'. Second, they were content-free.	Chi, M.T.H., Siler, S.A., Jeong, H., Yamauchi, T., & Hausmann, R.G. (2001), p. 508. See also Alexander, Hardman, and Hardman (2017).
3	Teacher poses questions asking students to reason, justify, analyse, and evaluate	The core qualitative difference evident among the interactions was how opportunities for student involvement in the conversations were opened up or closed down and how this led to different forms of student engagement. When teachers provided feedback in the form of questions that asked students to reason, justify, analyse, and evaluate their learning, they were inviting students into a dialogue. Some students then engaged	Adie, L., Kleij, F., & Cumming, J. (2018), p. 720. See also Ruiz-Primo (2007).
4	Teacher provides information that invites the students to continue their line of thinking	in self-analysis of their responses, and provided feedback to the teacher on strategies that worked best for them. In general, a scaffolding move is a kind of guided prompting that pushes the student a little further along the same line of thinking, rather than telling the student some new information, giving direct feedback on a student's response, or raising a new question or a new issue that is unrelated to the student's reasoning	Chi, M.T.H., Siler, S.A., Jeong, H., Yamauchi, T., & Hausmann, R.G. (2001), p. 490. See also Ruiz-Primo (2007).
5	Teacher establishes expectations for children's thinking and participation	… creating a context for argument in classrooms requires that teachers establish expectations for children's thinking and participation while they explain their solutions to others.	Bragg, L.A., Herbert, S., Loong, E.Y.-K., Vale, C., & Widjaja, W. (2016), p. 529. See also Webb et al. (2017).
6	Shared understanding of criteria and standards while evaluation is initiated mostly by the teacher	The development of students' understanding of the criteria and standards of performance is critical for students' informed contribution to feedback dialogues and to their development as self-assessors who are active agents and self-regulators of their learning.	Adie, L., Kleij, F., & Cumming, J. (2018), p. 721.
7	Students express their needs and areas for improvement	We found that the more teachers questioned and asked students for their feedback, or to identify areas where they required help, the more willing and articulate students were in expressing their strengths and areas for improvement.	Adie, L., Kleij, F., & Cumming, J. (2018), p. 720.
8	Students express personal agency, such as 'We noticed … ', 'I thought … ', 'I didn't understand … ', 'We decided … ', and 'We realised that … '.	Important reasoning words also included recount clauses and phrases that expressed personal agency, such as 'We noticed … ', 'I thought … ', 'I didn't understand … ', 'We decided … ', and 'We realised that … '.	Bragg, L.A., Herbert, S., Loong, E.Y.-K., Vale, C., & Widjaja, W. (2016), p. 527.
9	Students self-explain understanding and answer deeper questions	The results showed that students who were prompted to self-explain learned with greater understanding (were more able to	Chi, M.T.H., Siler, S.A., Jeong, H., Yamauchi, T., & Hausmann, R.G. (2001), p. 479. See also

(continued on next page)

Table 3 (*continued*)

Number	Category	Statements from sources	Source
		answer deeper questions) than a control group of students who were permitted to read the passage twice but not prompted to self-explain.	Ruiz-Primo (2007).
10	Students respond to guidance with comments and show understanding	*As listeners, students were also expected to voice their disagreement and to provide reasons for disagreeing. **Elicited responses: Basically, a scaffolding type of dialogue would be both constructive and interactive, in the sense that guidance would be provided by the tutors, and the students' would respond to the guidance with comments that followed up on what the tutors said. **This means that the dialogues were not interactive. The students did not participate in the explanation, to show that they understood it or to interject queries.	*Bragg, L.A., Herbert, S., Loong, E.Y.-K., Vale, C., & Widjaja, W. (2016), p. 529. ** Chi, M.T.H., Siler, S.A., Jeong, H., Yamauchi, T., & Hausmann, R.G. (2001), p. 504.
11	Students formulate their own in-depth thoughts and steps	This is a deep follow-up, which is an elaborative inference that extends what the tutor said …	Chi, M.T.H., Siler, S.A., Jeong, H., Yamauchi, T., & Hausmann, R.G. (2001), p. 491.
12	Student–student interaction is apparent	However, student–student interaction could, in principle, be increased further via activities that do not involve teachers.	Howe, C., & Abedin, M. (2013), p. 335.
13	Students self-analyse and reflect on their learning	*Some students then engaged in self-analysis of their responses and provided feedback to the teacher on strategies that worked best for them. **Reflection consisted of comprehension monitoring statements that were made either in response to tutors' comprehension gauging questions (CGQs), or to some other tutoring moves.	*Adie, L., Kleij, F., & Cumming, J. (2018), p. 720. ** Chi, M.T.H., Siler, S.A., Jeong, H., Yamauchi, T., & Hausmann, R.G. (2001), p. 491.

activity (Ruiz-Primo & Brookhart, 2018) and helps students understand the characteristics of good work. This scale consists of four items. A sample item is 'My students have to assess themselves in detail.' Answers were reported using a Likert scale ranging from 1 (do not agree at all) to 6 (fully agree). The Cronbach's α of the scale was 0.79.

Formative feedback from the teacher (student questionnaire). This scale mirrors the teacher's scale, but we were not able to use all the intended items in the reliability test. In the end, this scale, which measured students' perceptions of formative feedback, consisted of five items. A sample item is 'My teacher asks me to explain my way of solving the problem.' Answers were reported using a Likert scale ranging from 1 (do not agree at all) to 6 (fully agree). The Cronbach's α of the scale was 0.76.

Self-regulation during word-problem tasks (student questionnaire). The items measuring self-regulation were derived from the theory of self-regulated learning by Zimmerman (2002) and adapted from a study by Berger and Karabenick (2011). Self-regulation strategies are mostly subject specific (Boekaerts & Corno, 2005). The items enquired about students' use of self-regulation while solving word problems because not all were familiar with the term 'mathematical reasoning' at T1. A sample item for this assessment is 'When solving word problems, I plan my approach carefully.' Answers were reported on a Likert scale ranging from 1 (do not agree at all) to 6 (fully agree). The Cronbach's α of the scale was 0.77.

Self-efficacy in explaining solutions (student questionnaire). The three items on self-efficacy were newly constructed in line with a study by Pintrich and De Groot (1990). We focused on asking whether students were able to explain solutions because this was understandable for all students, which was not yet the case for the question about the ability to argue mathematically at T1. Being able to explain a concept is connected to achievement (Webb et al., 2017). A sample item is 'It is easy for me to explain why I calculate exactly this way.' Answers were reported using a Likert scale ranging from 1 (do not agree at all) to 6 (fully agree). The Cronbach's α was 0.68.

2.4.3. *Mathematical reasoning test*

All the items measuring mathematical reasoning were adapted from other standards-based tests or were developed by the project's team. The items were aligned with the basic competences of Switzerland (Swiss Conference of Cantonal Ministers of Education (EDK), 2011). As the items were open-ended, we expected a student to work on the 10 items for approximately 35–40 min during each session. All the responses were rated on a scale with four levels of competence, based on a rubric and a detailed manual with a description for each level (see sample items and manual excerpts in the Appendix).

The project's procedure and reliability were examined in a pilot project (Smit & Birri, 2014). Satisfactory inter-rater reliability was reached within each rating team: Kappa >0.70. The complete test battery consisted of 14 items, which were distributed in two testlets of 10 items each. Six items were used repeatedly to function as anchors for the item response theory (IRT) calibration. We applied the weighted least square mean and variance adjusted *(WLSMV)* estimator for ordinal data and used the graded response model of Mplus 8.4 (Muthén & Muthén, 2017). Upon completing the IRT analyses, final person measures, based on Bayesian plausible values (Von Davier, Gonzalez, & Mislevy, 2009), were computed for the reasoning tests. Plausible values were calculated for people with missing data. The factor score for mathematical reasoning at T1 was fixed at 0 (Var = 1), and the plausible values for the person parameters were calculated from a standard normal distribution. Hence, at T1 the mean was 0, and the SD was 0.81 at both measurement points. At T2, the mean was higher, $M = 0.60$.

2.5. *Data analysis*

2.5.1. *Analysis of video data using the many-facet Rasch measurement*

To examine our ratings, we identified the relevant lesson sequences and episodes of precise feedback (Ruiz-Primo & Min, 2013). Two coders viewed the recorded lessons of the 44 teachers and identified situations in which the feedback from the teachers helped students move from their current state of understanding to the next step toward mastery of the goal' (Ruiz-Primo & Brookhart, 2018, p. 50). Teachers were required to ask about and learn what students were thinking, and while doing this, they had the learning goals in mind and provided feedback

accordingly. Furthermore, 'the student needed to understand and use it [the feedback]' (Ruiz-Primo & Brookhart, 2018, p. 52). The two raters achieved an inter-coder reliability of 73–83%, or a Kappa of 0.64–0.71, for the basic coding.

The two raters coded each of the 13 categories in each selected lesson sequence (see Fig. 2, brick-red colour) by recording whether a behaviour was observed or not (0/1). Please note that the first category (number 0) was only to select the appropriate sequences. An overall rating was assigned to all the sequences in the lesson and for each category using a 3-point scale: 0 = not observed at all, 1 = sometimes observed, and 2 = often observed. Midway through the rating procedure, after rating half of the material, we conducted a preliminary check of the ratings' reliability using a many-facet Rasch analysis, which yielded the internal consistency of the rating model (see the Results section). Based on these results, we clarified the rating teams' understanding of the specific categories and continued with the remaining video lessons.

In the next step, a many-facet Rasch measurement analysis was conducted on the two rated sets of the 44 video lessons and 13 rated items using FACETS software, version 3.83.6 (Linacre, 2021). The Rasch model employs the principles of interval measurement to measure data objectively, which involves taking raw ordinal scores and performing a series of logarithmic transformations to produce data that support linearity. The many-facet Rasch measurement model is an extension of the original Rasch measurement model, as the analysis goes beyond person ability and item difficulty to measure other factors that interact in a testing situation. The model includes multiple facets (e.g. raters and occasions), so errors or variance based on different ratings are included in the model's calculations, such that differences could be detected in the levels of severity exercised by the raters when they rated a teacher's lessons.

2.5.2. Questionnaires

A multi-level regression analysis was used in this study because the units of analysis included teachers and students nested within classrooms. This procedure assumes that teachers influence students, and individual students, in turn, influence the properties of the class. Consequently, variables may be defined at the student and class/teacher levels. We applied a random intercept model.

The questionnaire that students completed on their perceptions of feedback quality and self-concept contained missing values (between 2.4% and 8.3%). Given that these missing values were not due to the study design, we assumed they occurred randomly. Thus, we initially applied the full-information maximum-likelihood procedure as a model-based treatment of missing data (Enders, 2010).

3. Results

3.1. Ratings of the observed interaction dialogue sequences

3.1.1. Statistical results and description of the wright map

First, we present the statistical results of the video analysis. As explained, the data derived from the rating process underwent a many-facet Rasch analysis. Fig. 3 presents the Wright map of the three facets: teacher (or class), rater, and category (item). The first column shows the linear, equal-interval logit scale, upon which all facets of the analysis are positioned, creating a single frame of reference for comparisons within and between the facets. This scale was revised to make it more readable by transforming the mean from 0 to 50 and the scale from 1 to 10 units per logit. The second column consists of the performance measures for the 44 teachers. Each star represents a teacher, and the teachers are ordered from highest quality of interaction dialogue (at the top of the column) to lowest quality (at the bottom). The third column displays the severity of the ratings by the two raters (A and B). It can be seen that they are on the same level, showing no difference in leniency. The fourth column displays the 13 categories from the rating manual and their difficulty. The more difficult categories (items) appear on the top of the figure and the easier ones at the bottom. It is apparent that the range of the ratings of the teachers' performance was smaller than that of the items' difficulty. This finding indicates that some of the items (5, 6, 11, and 13) were too difficult (i.e. out-of-range) for the teachers or could not be observed, although they were expected theoretically. However, two items at the bottom of Fig. 3 (8 and 4) were too easy for all the teachers. The fifth column shows the 3-point rating scale. A horizontal line across a column indicates the point at which the likelihood of a teacher receiving the next higher rating begins to exceed the likelihood of that teacher receiving the next lower rating. This line illustrates the threshold between the two rating-scale categories. It shows that none of the teachers scored an overall rating of 2 across all items, although some teachers reached a rating of 2 on some items, which allowed reliable calculations using the 3-point scale. The fit statistics of the many-facet Rasch measurement model consisted of infit and outfit values for each teacher and each category. In our case, the infit and outfit measures were between 0.5 and 1.5, which is within the range of well-fitting models (Wright & Linacre, 1994). Furthermore, rater reliability was 1.00, item reliability was 0.99, and person reliability was 0.74, which are considered good values for reliability and are comparable to Cronbach's alpha.

3.1.2. Description of the item difficulties

Among the easiest and most frequently observed' items were teachers '4 providing inviting information', which is a type of scaffolding that pushes the student further along the same line of thinking, and '2 providing open-ended and mostly non-specific prompts', such as 'Where are you?' and 'What have you been doing?' (Chi et al., 2001). When students were part of the interaction, they mostly '8 expressed agency', meaning they used phrases such as 'We noticed' and 'We had trouble understanding (something)', or they '10 responded to the teacher's guidance' with comments expressing their understanding (Bragg et al., 2016).

The next easiest and most frequently observed items were '1 unobtrusive feedback', '9 student explanations and expressions of deeper thinking'. Such observations included situations in which the teachers encouraged further dialogue and thought among the students rather than focusing on correcting mistakes, and asked students to self-explain what and how they understood something in a constructive way. The items of medium difficulty included teacher inviting students to identify areas where they '7 required help' and teacher '3 posing questions' asking students to reason, justify, analyse, and evaluate their learning. Other items of medium frequency were '12 student-student activities' in which interactions took place without the teacher (Howe & Abedin, 2013).

The items that were less frequent and more difficult were '6 providing criteria and standards for good reasoning', which would have been helpful for self-regulated learning (Adie et al., 2018). The most difficult and rarely observed items were teacher '5 asking students to explain and discuss their solutions with others' and teacher '11 encouraging students to formulate their own deeper connections with the steps of the process outlined by the teacher'. Finally, an item that was almost never observed was '13 student engagement in self-analysis of their strategies and reasoning steps'.

3.2. Model of interactive dialogue and student reasoning achievement

In the second step, we combined the teachers' performance data regarding interactive dialogue quality with the students' and teachers' questionnaire data to explore further associations predicted in our research model (Fig. 1). Given that many of the students were not acquainted with the tasks or learning environment provided in our study, we did not examine the longitudinal data but applied data from the post-questionnaires only when all the students' and teachers' answers pertained to our project and the content of our subject. This method was seen as the most reliable and valid approach for achieving a

Rater 1		Categories	0	1	2	3	4	5	6	7	8	9	10	11	12	13	
T = Teacher; S = Student			T-S	T	T	T	T	T	T	S	S	S	S	S	S	S	Sum
Name	**Start**	**End**	Interaction has several turns	Teacher provide only wanted feedback	Teacher prompts are open-ended and content-free	Teacher poses questions asking students to reason, justify, analyse and evaluate	Teacher provides information that invites the students to continue their line of thinking	Teacher establishes expectations for children's thinking and participation	Shared understanding of criteria and standards while evaluation is initiated mostly by the teacher.	Students express their needs and areas for improvement	Students express personal agency, such as We noticed..., I thought..., I didn't understand..., We decided..., and We realised that....	Students self-explain understanding, answer deeper questions	Students respond to guidance with comments, shows understanding	Students formulate their own in-depth, further thought steps	Student-student interaction is apparent	Students self-analyse, reflect their learning	
21HHK	0:00:46,7	0:01:19,5															0
21HHK	0:01:22,6	0:01:49,9															0
21HHK	**0:02:07,2**	**0:03:26,3**															0
21HHK	0:03:45,2	0:03:49,5															0
21HHK	0:04:17,8	0:04:23,8															0
21HHK	**0:04:25,8**	**0:05:13,9**	1				1				1	1					4
21HHK	0:05:20,9	0:05:53,2															0
21HHK	**0:06:26,4**	**0:08:53,5**	1		1		1				1		1				5
21HHK	**0:09:03,5**	**0:09:32,7**	1		1		1				1						4
21HHK	0:09:40,9	0:09:42,5															0
21HHK	0:09:57,5	0:10:04,6															0
21HHK	**0:10:13,3**	**0:11:22,6**	1	1	1		1			1	1	1	1				8
21HHK	0:11:39,6	0:11:49,0															0
21HHK	0:12:10,7	0:12:22,0															0
21HHK	**0:12:35,3**	**0:13:28,7**	1		1		1				1						4
21HHK	0:14:01,7	0:14:16,3	1	1	1		1			1							5
21HHK	**0:14:28,6**	**0:16:29,7**	1		1		1				1	1	1				6
21HHK	0:16:59,2	0:17:27,2	1	1	1		1			1	1	1					7
21HHK	0:17:34,9	0:17:35,7															0
21HHK	**0:17:45,8**	**0:18:48,0**	1	1	1		1			1	1	1	1				8
21HHK	0:18:52,0	0:25:10,0	1		1	1		1	1		1	1	1	1	1	1	11
21HHK	0:25:39,6	0:28:42,4	1			1		1	1		1	1	1		1	1	9
		Sum	11	4	9	2	9	2	2	4	10	7	6	1	2	2	71
		Rating		1	2	1	2	1	1	1	2	1	1	1	1	1	16

Fig. 2. Coding sheet for rater 1, with a final rating for teacher 21HHK; 3-point scale: 0 = not observed at all, 1 = sometimes observed, and 2 = often observed.

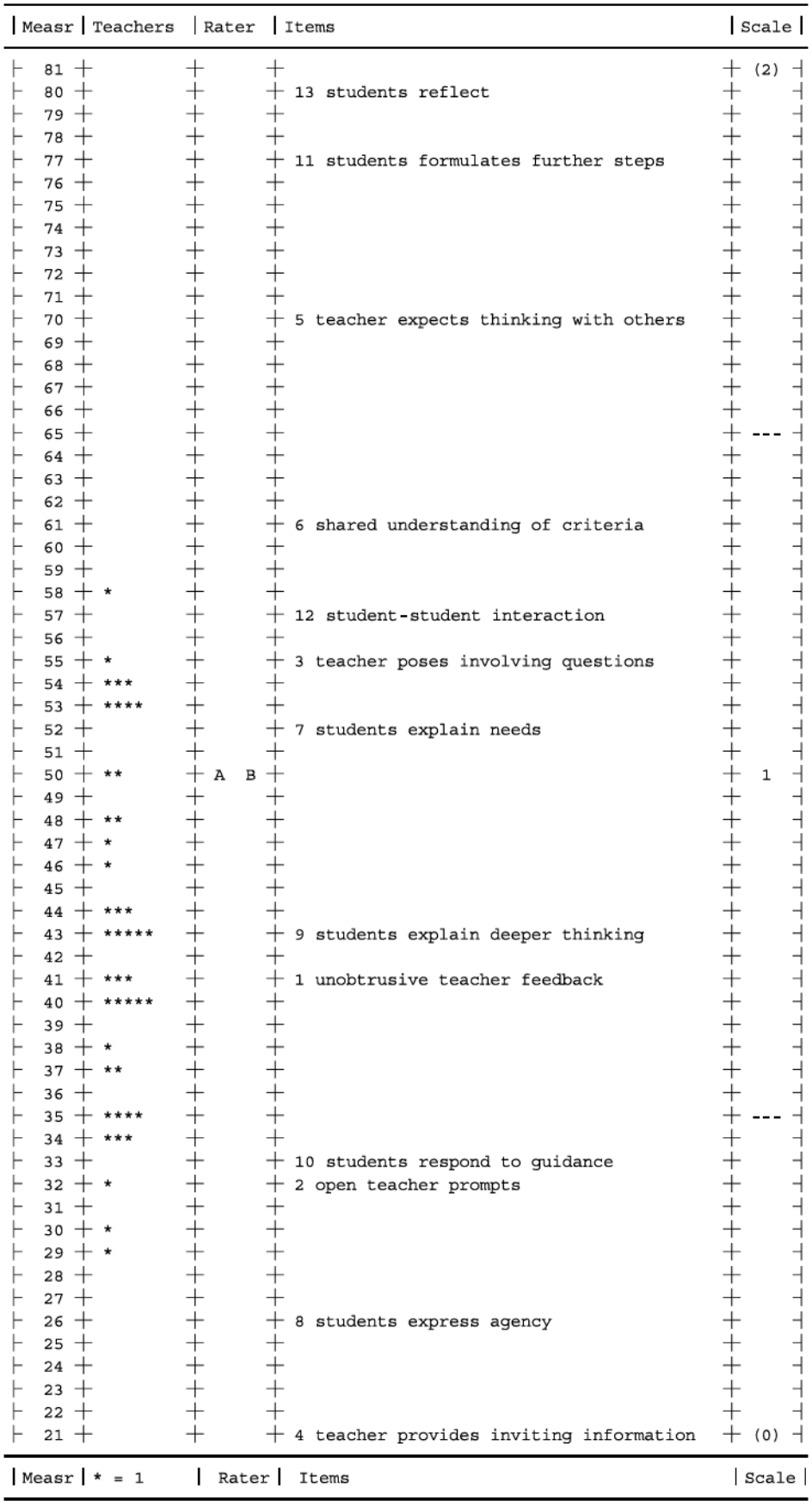

Fig. 3. Wright map of teachers' classroom interaction dialogue performance, inter-rater agreement between A and B, and item difficulty (Measr = measure; mean = 50; scale = 10; $N = 44$).

common understanding of the constructs within our study.

Descriptive statistics and correlations for within-person (i.e. student) and between-person (i.e. class) variables are presented in Table 4. On average, the students reported fairly good levels of mathematical explaining skills ($M = 3.63$ on a 6-point scale). However, students' scores were found to be quite variable ($SD = 0.97$), indicating the full use of the scale. The mean of the students' perceived self-regulation skills for solving word problems was similar to the mean of their perceived level of mathematical explaining skills ($M = 3.88$), but the variance was lower ($SD = 0.67$). Perceptions of teachers' formative feedback were also fairly high ($M = 3.63$, $SD = 0.88$).

Zero-order correlations at the student level showed a positive relationship between mathematical reasoning competence and perceived explaining skills ($r = 0.28$). Explaining skills and formative feedback were positively correlated, but less so ($r = 0.15$). The strongest correlations were found between self-regulation and explaining skills ($r = 0.41$) and self-regulation and formative feedback ($r = 0.35$). At the class level, we again found medium correlations between reasoning competence and perceived explaining skills ($r = 0.27$) and between explaining skills and formative feedback ($r = 0.30$). However, none of these two correlations were significant due to high standard errors. The correlation of self-regulation with formative feedback was 0.38, and with explaining skills it was 0.73. Interactive dialogue showed a significant relationship with explaining skills ($r = 0.43$) and student-student feedback ($r = 0.45$), a variable from the teachers' questionnaire.

Student-student feedback was positively correlated with students' perceived formative feedback practices ($r = 0.32$). The intra-correlation coefficients (ICCs) showed that approximately 14% of the variance in mathematical reasoning was due to the class. Hence, it is reasonable to explore explanations for separate differences at the student and classroom levels. The ICCs for formative feedback were substantially high (18%), whereas those for self-regulation were lower (11%). The ICC for self-efficacy for explaining indicated little variance between the classes (4%).

Based on the results of a previous study (Smit et al., 2022), we knew that students' reasoning test results were predicted by formative feedback and mediated by students' self-efficacy beliefs. Therefore, we initiated the construction of our model in this study with a similar model (see Fig. 1). We added students' perceived self-regulation skills for solving word problems, similar to another previous study (Smit et al., 2017) which showed that formative feedback predicted self-regulation, mediated by peer- and self-assessment practices. In a subsequent step of the model's construction, we tested the influence of variables in the student context at the individual level. Among these variables, 'book topics at home' (see Verhoeven & van Elsäcker, 2016) was a significant predictor of explaining skills ($\beta = 0.14$) and mathematical reasoning ($\beta = 0.11$). Next, we added the ratings of the interactive dialogues from the video analysis at the between level. Because we intended to include variables for the teachers' perception of formative feedback practices, we constructed a model with three scales: diagnostic skills, formative feedback, and self- and peer-assessment. However, none of the models produced satisfying fit models, so we switched to single items and chose those that aligned with items from the rating scale for interactive dialogue. Only one of the teachers' items appeared to be related to interactive dialogue: student-student feedback practices. We compared a few models with differing path directions but identical variables. Based on the deviance information criteria, we selected a model with all paths on the class level, starting from the variable 'interactive dialogue', which was temporally prior to the other variables.

The final model (Fig. 4) had well-fitting fit indices. We calculated the model using Mplus 8.4 with a maximum likelihood estimation: $\chi^2/df = 0.90$ (4.88/10); CFI/TLI = 1.00; RMSEA = 0.000; SRMRw = 0.01; and SRMRb = 0.08.

In our first hypothesis, we expected the quality of teachers' interactive dialogue to have an indirect effect on students' competence in mathematical reasoning via the students' perceived self-efficacy for explaining. This hypothesis, as a whole, was not supported. The paths presented in Fig. 4 and Table 5 show, at the class level, a significant effect of interactive dialogue on the mediator 'self-efficacy student explaining' ($\beta = 0.30$), which is a small to medium effect size. However, there was no significant effect of the mediator 'self-efficacy for explaining' on mathematical reasoning ($\beta = -0.12$).

Regarding our second hypothesis we expected the quality of teachers' interaction dialogue to have an indirect effect on students' competence in mathematical argumentation via the students' perceived self-regulation skills for word problems. At the class level, no significant effect of interactive dialogue on the mediator 'self-regulation' ($\beta = 0.06$) existed, but a medium effect ($\beta = 0.45$) of the mediator 'self-regulation' on mathematical reasoning was found. However, this effect was not significant due to the high standard errors. The indirect effect of interactive dialogue on mathematical reasoning via self-efficacy for explaining was not significant ($\beta = 0.01$) either. In addition, the total indirect effect of the interactive dialogue on mathematical reasoning via all four paths was negligible ($\beta = 0.02$).

Our third hypothesis stated that the quality of teachers' interactive dialogue is related to students perception of the teachers' formative feedback practices and the teachers' use of peer feedback. This hypothesis was confirmed for peer feedback only. To test this hypothesis, we established a direct path from interactive dialogue to formative feedback and an indirect path via opportunities for student-student feedback in the model. Our results showed that the teachers whose interactive dialogue performance was considered to be at a high level indicated that their students were provided with opportunities to give each other comprehensive feedback more often ($\beta = 0.39$). However, the student-student feedback item from the teacher questionnaire was not a significant predictor of students' perceived formative feedback ($\beta = 0.18$), and the effect size was small. In addition, no significant direct effect of interactive dialogue on formative feedback was found ($\beta = -0.02$). Thus, the student, teacher, and expert views of supportive practices for mathematical reasoning in the classroom aligned only for

Table 4
Descriptive statistics and correlations between variables at the within-person and between-person levels of analysis.

Variable	*M*	*SD*	1	2	3	4	5	6	7	8	9
Within-person level											
1. Formative feedback	3.63	0.88	–	–	–	–	–	–	–	–	–
2. Self-regulation	3.88	0.67	0.35**	–	–	–	–	–	–	–	–
3. Self-efficacy explaining	3.63	0.97	0.15**	0.41**	–	–	–	–	–	–	–
4. Mathematical reasoning	0.60	0.81	−0.01	0.07	0.28**	–	–	–	–	–	–
Between-person level											
5. Formative feedback	–	–	–	–	–	–	–	–	–	–	–
6. Self-regulation	–	–	–	–	–	–	0.38*	–	–	–	–
7. Self-efficacy explaining	–	–	–	–	–	–	0.30	0.73**	–	–	–
8. Mathematical reasoning	–	–	–	–	–	–	0.14	0.22	0.27	–	–
9. Student-student feedback	3.81	1.72	–	–	–	–	0.32**	−0.01	0.11	0.04	–
10. Interactive dialogue[a]	42.92	7.57	–	–	–	–	0.25	−0.06	0.43**	0.06	0.45**

Note: $N_{within} = 804$; $N_{between} = 44$; Likert scale 1–6, the poles were: 6 = *absolutely agree*, and 1 = *absolutely disagree*; [a]*Rasch performance measure*; **$p < .01$, *$p < .05$.

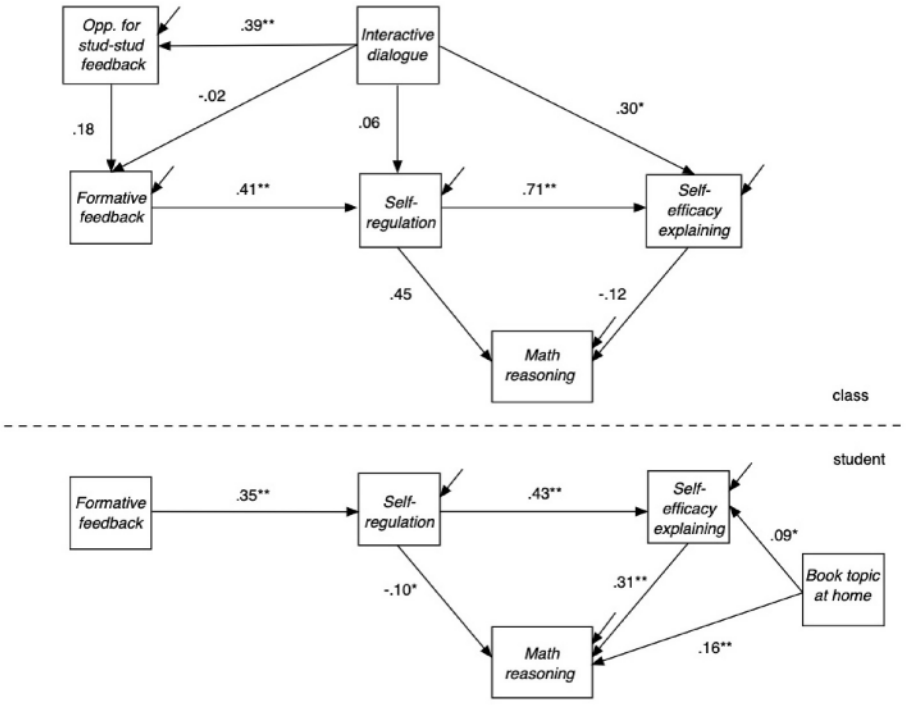

Fig. 4. Multi-level structural equation model of students' explaining skills and reasoning outcomes regressed on classroom interactive dialogue and formative feedback. $N_{within} = 804$; $N_{between} = 44$; $^{**}p < .01$, $^{*}p < .05$.

Table 5
Results of the multi-level models' predictions of students' mathematical reasoning outcomes.

Regression paths (student perceptions combined with teacher perception and expert rating)	Beta estimates	*SD*
Student Level		
Self-efficacy explaining → Math reasoning	0.31**	0.03
Self-regulation → Math reasoning	−0.10*	0.04
Book topics at home → Math reasoning	0.16**	0.04
Self-regulation → Self-efficacy explaining	0.43**	0.04
Book topics at home → Explaining skills	0.09*	0.04
Formative feedback → Self-efficacy explaining	0.35**	0.04
Teacher/Class level		
Self-efficacy explaining → Math reasoning	−0.12	0.47
Self-regulation → Math reasoning	0.45	0.44
Self-regulation → Self-efficacy explaining	0.71**	0.21
Interactive dialogue → Self-efficacy explaining	0.30*	0.13
Formative feedback → Self-regulation	0.41*	0.17
Interactive dialogue → Self-regulation	0.06	0.16
Interactive dialogue → Student-student feedback (teacher perception)	0.39**	0.11
Student-student feedback (teacher perception) → Formative feedback	0.18	0.20
Interactive dialogue → Formative feedback	−0.02	0.93

Note: $N_{within} = 804$; $N_{between} = 44$; $^{**}p < .01$, $^{*}p < .05$.

peer feedback, as perceived by the teacher and the interactive dialogue observed by the expert. In our lesson script, students were expected to give each other feedback.

Thus far, we have only discussed the classroom level; to answer the fourth hypothesis we also examined the student level. We expected teachers' formative feedback practices to have an indirect effect on students' competence in mathematical reasoning via the student's perceived mathematical explaining and self-regulation skills. This hypothesis was supported, but the results varied by student and classroom level. It appears that the paths from formative feedback to mathematical reasoning via self-efficacy for explaining and self-regulation reached significance only at the individual level only, and not at the classroom level. At the classroom level, formative feedback predicted self-regulation, with $\beta = 0.41$, and self-regulation predicted self-efficacy for explaining, with $\beta = 0.71$. Furthermore, self-regulation predicted mathematical reasoning, with $\beta = 0.45$, but the confidence interval was large, and significance was not reached. Explaining skills had no significant impact on mathematical reasoning, and the effect was small ($\beta = -0.12$). The two indirect paths from formative feedback to mathematical reasoning were not significant, and the total effect size was also small ($\beta = 0.15$).

At the student level, formative feedback was a significant predictor of reasoning competence, mediated by self-regulation and explaining skills. The fixed effects sizes were slightly smaller than those at the class level, but the standard deviations were smaller as well, and significance was achieved for all paths. The path from formative feedback to self-regulation showed a standardised regression effect of $\beta = 0.35$, similar to the path from self-regulation to explaining skills $\beta = 0.43$. Finally, explaining skills predicted reasoning competence, with $\beta = 0.31$, and self-regulation had a minor effect on reasoning competence, with $\beta = -0.10$. The total indirect effect of formative feedback on mathematical reasoning was not significant and very small ($\beta = 0.01$).

Overall, the model mostly helped to explain the variance at the between level in self-efficacy for explaining (62%), self-regulation (18%), and student-student feedback (15%). At the individual level, 20% of the variance in self-efficacy for explaining and 12% in self-regulation was explained. The model explained less well the variance in mathematical reasoning between classes (12%) and students (11%) – that is, the strength of the model lies in explaining class differences in students' efficacy beliefs, with respect to explaining one's own actions or the mistakes of others, which is also a goal of mathematical reasoning. Furthermore, although the model could not explain well the effect of interactive dialogue on class performance in mathematical reasoning, it was helpful for understanding differences in students' perceptions of feedback related to mathematical reasoning.

4. Discussion

The purpose of this study was twofold. First, we rated the quality of interactive dialogue within feedback episodes in lessons in which primary students worked on reasoning tasks. Using a rating system with 13 items, we found that most teachers showed skills that were typical of an effective interactive dialogue (Chi et al., 2001; Ruiz-Primo, 2011). This result does not mean that non-interactive dialogue was absent from the feedback episodes; we did not record it. Typical skills included starting dialogues with open-ended questions intended to assess the students' level of progress. Other moves included scaffolding that invited students to continue their line of thinking, along with feedback, which was used with restraint. As part of such scaffolding moves, short explanations were observed. Students mirrored these typical teacher feedback moves by expressing aspects of their own actions or self-explaining these actions. The student side of the dialogue included comments on the teachers' feedback or prompts. Less frequently students shared formulations of their needs and specific areas in which they needed help; instead, teachers identified areas for improvement most of the time. Deeper questions from the teachers also occurred requiring students to reason, justify, analyse, and evaluate their procedures, but not very often.

Although student-student activities were included in the design of our lessons, not many occurred during the teachers' interactive dialogue. To clarify this, group activities were present, but they were seldom accompanied by teachers' scaffolding episodes that involved more than one student. Many of the teachers possessed a rubric for reasoning, but surprisingly they rarely asked students to compare their reasoning solutions with the criteria that were provided. The teachers almost never formulated expectations of students to argue with each other or to think along the same line of thinking as the other students. We seldom observed deep follow-ups by students, as described in Chi et al. (2001) and Ruiz-Primo (2011), consisting of elaborative inferences extending teachers' statements or more detailed verbalisations of their reflections of what was and was not understood.

All in all, the observed feedback dialogues were mostly teacher led, with students merely responding to the teachers' moves, as in the studies by Chi et al. (2001) and Stovner and Klette (2022). Student-student interaction (Howe & Abedin, 2013) as part of the dialogic interaction was rare, even though it was suggested in our lesson plan. However, during all the lessons, interactive dialogue was present in the form of explaining, scaffolding, and giving feedback. These dialogues were often short-lived because the next student was already waiting for the teacher. Therefore, it was rather difficult for the learners to discuss their own thoughts with the teacher in depth. For lower-scoring pupils, teachers often resorted to modelling possible courses of action, perhaps because of time constraints or the students' reluctance or inability to initiate the process. This scenario may be considered being borderline interactive; Jones, Tanner, and Treadaway (2000) described it as 'funnelling', a less dynamic form of scaffolding, that is, guiding the student along predetermined steps. However, the risk exists in such scaffolding that the student will not be able to follow the teacher's train of thought and will not know what to do in the end (Munson, 2019).

Overall, it is notable that relatively few instances were observed of learners self-regulating their learning, be it by checking their solutions with given criteria or by reflection in general. We address this topic in the section below on the educational significance of our study.

In the second step of our study, we analysed the relationship of the data on interaction quality and students' reasoning competence as mediated by formative feedback, self-regulation skills, and self-efficacy beliefs. Our multi-level data – students nested in classes – were analysed using the data yielded by the research questions at the level of the individual student and the class. Overall, we could not confirm our hypothesis regarding the relationship between interactive dialogue and mathematical reasoning. However, the quality of the interactive dialogue was found to be of importance for students' self-efficacy beliefs for explaining. In our previous longitudinal study (Smit et al., 2022), we were able to show that the development of self-efficacy explains performance in mathematical reasoning over time.

Correlations were found between interactive dialogue and the factors related to acquiring mathematical reasoning skills. Most important, we found specific differences and parallels in the results for these factors at each level. The findings at the student level were consistent with our hypothesis that students' individual perceptions of formative assessment practices predict mathematical reasoning competence, mediated via self-regulation skills and self-efficacy beliefs, operationalised here as self-efficacy for explaining skills. We interpret this finding to mean that problem-solvers proficient in reasoning perceived their self-explaining and self-regulation skills as high and their teachers' formative feedback practices as beneficial for learning. Socioeconomic background also had an impact on explaining skills and reasoning competence, and other research has also shown that students from higher socioeconomic backgrounds have advantages in arithmetic tasks with open-ended task formats (Schwabe, Trendtel, & McElvany, 2019).

One difference between the student and class levels was that self-regulation skills had only a small direct effect on reasoning competence at the student level, but they had a medium direct effect at the class level. This finding indicates that classes with high self-regulation skills have better reasoning competence and higher self-efficacy for explaining; however, at the class level, no association was found between reasoning competence and self-efficacy for explaining. We considered how to interpret this finding – the ICC values suggest that the teachers worked differently regarding fostering students' self-regulation skills, and this factor seems to have an effect on their mathematical reasoning competence. However, for the individual student, self-efficacy beliefs related to word-problems are more relevant than self-regulation skills with respect to reasoning. Hence, teachers should have class-level goals in mind for promoting self-regulation skills and support individual students in their self-efficacy during interactive dialogue. Self-efficacy is best enhanced by self-referenced feedback and not by comparing oneself to others (Hattie & Gan, 2011).

4.1. Limitations and future directions

In our study, we distributed a prepared lesson plan to all the teachers, but each teacher interpreted and implemented it personally. In addition, external circumstances in the school required adjustments to our lesson plans. For example, many of the lower-scoring students required additional explanations from the teacher, causing the classes to move slowly. Regarding this change, no certainty exists that what occurred in these classes can be compared exactly. However, classroom interventions are, and should be, subject to unique adaptations by individual teachers according to the exigencies of their own curricula, values, and beliefs (Randi & Corno, 1997). In general, such adaptations to the local classroom culture benefit students' learning (Squire, MaKinster, Barnett, Luehmann, & Barab, 2003), but in our study, teachers' lack of understanding of the project's goals could have prevented the implementation intended, making interpretation of the results difficult. Therefore, it is possible that the teacher performances in the videos are not accurate representations of their competence in the field of interactive dialogue management. We recorded only one lesson on video, so it is possible that the teachers exhibited the unobserved behaviour in another lesson. However, it is also possible that they attempted to show their best teaching skills during the video recordings.

Questionnaires in general are limited by their inadequate measures of abstract constructs, such as teaching mathematical reasoning in the classroom (Johnson & Christensen, 2012). To ensure adequate measurement of complex constructs, we used summated rating scales and multiple methods whenever possible. We believe that the study's video analysis and two types of questionnaires complemented each other and facilitated a more accurate interpretation of what was happening in the classroom.

Although we designed a structural equation model with plausible causes and effects, only the video data were collected at an earlier stage, and the teacher and student questionnaire data were collected at the same time. However, we followed the causal chain as shown in the 'cascade model' (Krauss et al., 2020), where it is believed that the teachers' behaviour implies subsequent student learning. Due to the cross-sectional design of the study, we cannot infer effects over time, but we consider our approach to clarifying the relationship between teacher behaviour and student learning to be meaningful and significant. A longitudinal design using all the variables would be complex and challenging to analyse, and it would require an understanding of mathematical reasoning among all students at the beginning of the project. As not all the teachers applied such tasks in their math instruction before our project this problem remains to be solved.

Another limitation is that we did not include motivational-affective aspects of learning mathematical reasoning in our study. Chi et al. (2001) concluded from their study that an interactive style of dialogue could be more motivating for students and lead to greater enjoyment of learning. Boekaerts and Corno (2005) found that favourable appraisals of tasks and opportunities for learning (e.g. relevance or interest) motivate students and promote perseverance, whereas unfavourable appraisals (e.g. difficult, irrelevant, or stressful) lead to a focus on well-being. Consequently, students anticipating negative feelings try to avoid failure by dawdling, or waiting for the teacher's explanation of the task instead of trying to find their own solutions. The teachers participating in our study told us that towards the end of the implementation phase, some of their students lost interest and did not persevere. Teachers can use praise - although not recommended by Hattie and Timperley (2007) - to encourage student engagement, thereby positively affecting students' self-efficacy and perseverance in a task (Jain, Bruce, Stellern, & Srivastava, 2007; Medway & Venino, 1982). One should adopt a more differentiated view of Hattie's statement that feedback on the self-level is not beneficial for learning (Hattie & Timperley, 2007). Perhaps the problem is that teachers provide too much feedback at the self-level without adding feedback on the task or the process at the same time (Hattie & Gan, 2011).

Although we did not separately analyse the mathematical content of the dialogues, parallel analysis of the mathematical content of the videotaped teacher-student conversation is underway. We intend to analyse the corresponding data in combination at a later date. We anticipate that it will be interesting to clarify whether good interaction dialogue is related to high amounts of feedback regarding the argumentation or whether it is more about finding an approach to solve the problem.

Following the research of Fyfe and Brown (2018), it would be worthwhile to distinguish teachers' interaction dialogue in relation to the students' competence levels, as competent students need little or no feedback and less competent students need not only feedback but also modelling by the teacher. It could be that our videos included mostly interactions with less-competent students. This process would require the teachers of each class to identify students according to competence level in each coded video sequence, so it would be important to consider how we could address this methodological challenge in future research.

4.2. Theoretical and educational significance

To what extent can our data be attributed to research in the field of interactive dialogue? As Alexander (2018) mentions, there is an overlap between formative assessment or feedback and interactive dialogue. Our approach of interactive dialogue as one-to-one feedback episodes is only part of the possible interactive settings (Alexander, 2018). In this respect, the principles Alexander uses to define interactive dialogue cannot be fully mapped with our data. For example, collectivity in the classroom is not perceptible in our video recordings. The camera focuses the teacher during her/his feedback activities, discussions between students without a teacher are not seen as often and are limited to two peers. Collective dialogues in the class were not foreseen in the recorded lesson. Cumulative aspects are sometimes visible in the video recordings, especially when teacher and student(s) are engaged in productive, longer feedback conversations. We discuss this below. Often, however, the teacher is more of a leader in the conversation, so that dialogue is limited to one-sided contributions. This is reflected in lower Rasch performance scores.

In the following, we examine some of the categories of the video rating manual and relate our findings to the study's theoretical framework and classroom practice. We start with the frequently occurring 'teacher explanations' and proceed to the rarely occurring 'student reflections' and 'references to self-regulation'. Research is relatively consistent in reporting the lack of effectiveness of feedback in the form of explanations by teachers during interactive dialogues. We would like to put this finding into perspective. Contrary to the study by Chi et al. (2001), in our study, explanations as part of mathematical reasoning were probably useful for initiating the first approach to a possible line of argumentation, at least for a portion of the students. These explanations could have included clarifying a word problem, a question, or the meaning of a certain term (e.g. distinguishing a digit from a number). This is consistent with the results of Fyfe, Rittle-Johnson, and DeCaro (2012), who found that children with little prior knowledge needed more feedback that provided conceptual knowledge. Within the framework of Drageset (2015), this teacher feedback approach proceeds through the phases of explanations based on the pattern of teacher-led responses and progressive actions. Our study's classroom videos suggest that it might be necessary for teachers to provide some students with a model of what a first calculation could look like, as part of a line of argumentation. However, according to Drageset (2015), to reach the full potential of an interactive dialogue, the teacher should also use questions to gain further insights into students' thinking without providing an answer during the initial phase. During the latter, more cooperative problem-solving phase of scaffolding, the teacher should interfere only when the student is stuck or the line of argumentation is unfinished, as this might foster self-regulation skills (Kramarski & Mevarech, 2003).

Reflection, as part of metacognitive discourse (Shilo & Kramarski, 2019), is crucial for self-regulation (Butler & Winne, 1995), but in our study it rarely occurred in the observed lesson sequences. Wheatley (1992) views reflection as central to developing skills to solve novel mathematics problems and construct new knowledge. In the study by Chi et al. (2001), students' reflections were the most effective behaviour, and we think that a lack of reflection is a finding that deserves greater attention from mathematics teachers. We also suggest that teachers improve their methods for fostering self-regulation skills for mathematics by encouraging students to make extended rather than brief responses to the teacher's initiations, to supplement their responses with underlying reasoning, and to discuss other students' contributions rather than rely upon the teacher's feedback (Adie et al., 2018; Stovner & Klette, 2022). For students who struggle to understand the language used in their feedback, educators could provide more opportunities for them to improve their assessment literacy to understand what is expected of them (e.g. apply the grading criteria to their own work) (Winstone et al., 2017). Similar to the results of Adie et al. (2018), few dialogic interactions in our study involved negotiations of the criteria or achievement standards and how they could be used to understand the quality of students' reasoning. Although some teachers possessed a rubric, we rarely observed teachers or students checking that they shared an understanding of the criteria or the required standard for a response. Throughout the feedback interactions, few teachers questioned students about their understanding of and responses to the rubric criteria or self-formulated ones. Developing students' understanding of criteria and performance standards is critical to their contributions to interactive dialogues in a manner that promotes their development as self-assessors, which in turn, is part of their self-regulated learning (Meusen-Beekman et al., 2016; Perry, 1998). The suggested cooperative methods of learning in our lesson plans supports students'

self-assessments as an important part of their self-regulation (Kramarski & Mevarech, 2003).

Fostering self-regulation skills in students could relieve teachers of unnecessary tasks, allowing them more time to talk with individual students, which seemed to have been a problem in the observed lessons. For the lower-scoring pupils, such discussions could be interactive dialogues with respect to closing gaps in knowledge and understanding, and for the higher-scoring pupils, they could involve discussing individual approaches to reasoning tasks (Fyfe et al., 2012). As students have different needs, the right balance between guidance and autonomy is an individual difference. On the one hand, problem-solving by students with low mathematical skills can lead to an overload of their cognitive capabilities, which can be lessened with guidance from the teacher (Fyfe & Brown, 2018). On the other hand, the teacher should have the courage not to give unnecessary instructions to students (Adie et al., 2018) and restrain themselves from doing so when working on a problem (Fyfe & Brown, 2018).

The social organisation of the lessons in our study proved useful for learning to reason and to argue. Student-student-feedback seems to be a relevant feature of this organisation. As Wyatt-Smith and Adie (2019) noted, the presence of other learners provides students with calibrations of their own progress, helping them to identify their strengths and weaknesses initiate efforts for improvement. Teachers should prepare students for such cooperative sessions of interactive dialogue by helping them develop their discourse competence, which plays an important role in the learning of mathematics (Erath, Prediger, Quasthoff, & Heller, 2018). Discourse competence also includes that learners have an adequate vocabulary. Within mathematical reasoning, the acquisition of words, such as 'because', 'therefore', and 'for example' provide students with the necessary language to participate in the discursive and epistemic process of knowledge creation (Erath, Ingram, Moschkovich, & Prediger, 2021).

Although our study did not generate new knowledge about the relevance of interactive dialogue within feedback episodes, student beliefs, and mathematical reasoning competence, it appears to be the first study to include all the variables together in one empirical model to explore how these factors interact between the teacher and the students and among individual students. Another strength of the present work is the integration of the different perspectives of teachers, students, and experts on teaching and learning mathematical reasoning. Finally, the study identifies which factors teachers should pay attention to when they engage students in reasoning tasks. It was found that teachers differ in their interactive dialogue and feedback competence in terms of promoting students' self-efficacy in explaining solutions. In addition, it was found that students with high self-regulation skills have higher mathematical reasoning competence and that the perceived feedback quality of the teacher plays a role in this.

Funding

The present project was funded by the Swiss National Science Foundation (Project no. 100019_179230).

Author statement

Robbert Smit: Writing – Original draft preparation, reviewing and editing, Methodology, Formal analysis, Conceptualization, Funding acquisition. **Kurt Hess:** Conceptualization, Funding acquisition, Writing - Review & Editing. **Alexandra Taras:** Formal analysis, Writing - Review & Editing. **Patricia Bachmann:** Validation; Formal analysis; Project administration, Investigation. **Heidi Dober:** Validation; Formal analysis; Investigation.

All authors have read and agreed to the published version of the manuscript.

Acknowledgements

The present project was funded by the Swiss National Science Foundation (project no. 100019_179230).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.learninstruc.2023.101777.

References

Adie, L., Kleij, F., & Cumming, J. (2018). The development and application of coding frameworks to explore dialogic feedback interactions and self-regulated learning. *British Educational Research Journal, 44*(4), 704–723. https://doi.org/10.1002/berj.3463

Alexander, R. (2018). Developing dialogic teaching: Genesis, process, trial. *Research Papers in Education, 33*(5), 561–598. https://doi.org/10.1080/02671522.2018.148114

Alexander, R., Hardman, F. C., & Hardman, J. (2017). *Changing talk, changing thinking: Interim report from the in-house evaluation of the CPRT/UoY Dialogic Teaching Project.* University of York and Cambridge Primary Review Trust. https://eprints.whiterose.ac.uk/151061/.

Ball, D. L., & Bass, H. (2000). Making believe: The collective construction of public mathematical knowledge in the elementary classroom. In D. C. Philipps (Ed.), *Constructivism in education: Opinions and second opinions on controversial issues. Yearbook of the national society for the study of education* (pp. 193–224). University of Chicago Press.

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review, 84*(2), 191–215. https://doi.org/10.1037/0033-295X.84.2.191

Bandura, A. (1997). *Self-efficacy: The exercise of control.* W.H. Freeman and Company.

Berger, J.-L., & Karabenick, S. A. (2011). Motivation and students' use of learning strategies: Evidence of unidirectional effects in mathematics classrooms. *Learning and Instruction, 21*(3), 416–428. https://doi.org/10.1016/j.learninstruc.2010.06.002

Bezold, A. (2009). *Förderung von Argumentationskompetenzen durch selbstdifferenzierende Lernangebote. Eine Studie im Mathematikunterricht der Grundschule.* Kovac.

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability, 21*(1), 5–31. https://doi.org/10.1007/s11092-008-9068-5

Blum, W., & Kirsch, A. (1991). Preformal proving: Examples and reflections. *Educational Studies in Mathematics, 22*(2), 183–203. https://doi.org/10.2307/3482408

Boekaerts, M. (1999). Self-regulated learning: Where we are today. *International Journal of Evaluation and Research in Education, 31*, 445–457.

Boekaerts, M., & Corno, L. (2005). Self-regulation in the classroom: A perspective on assessment and intervention. *Applied Psychology: International Review, 54*(2), 199–231.

Bragg, L. A., Herbert, S., Loong, E. Y.-K., Vale, C., & Widjaja, W. (2016). Primary teachers notice the impact of language on children's mathematical reasoning. *Mathematics Education Research Journal, 28*(4), 523–544. https://doi.org/10.1007/s13394-016-0178-y

Brodie, K. (2010). *Teaching mathematical reasoning in secondary school classrooms.* Springer.

Brown, G. T. L., Harris, L. R., & Harnett, J. (2012). Teacher beliefs about feedback within an assessment for learning environment: Endorsement of improved learning over student well-being. *Teaching and Teacher Education, 28*(7), 968–978. https://doi.org/10.1016/j.tate.2012.05.003

Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research, 65*(3), 245–281.

Chi, M. T. H., Siler, S. A., Jeong, H., Yamauchi, T., & Hausmann, R. G. (2001). Learning from human tutoring. *Cognitive Science, 25*(4), 471–533. https://doi.org/10.1016/S0364-0213(01)00044-1

Cummins, D. D., Kintsch, W., Reusser, K., & Weimer, R. (1988). The role of understanding in solving word problems. *Cognitive Psychology, 20*(4), 405–438. https://doi.org/10.1016/0010-0285(88)90011-4

Davis, B. (1997). Listening for differences: An evolving conception of mathematics teaching. *Journal for Research in Mathematics Education, 28*(3), 355–376. https://doi.org/10.2307/749785

Drageset, O. G. (2014). Redirecting, progressing, and focusing actions—a framework for describing how teachers use students' comments to work with mathematics. *Educational Studies in Mathematics, 85*(2), 281–304. https://doi.org/10.1007/s10649-013-9515-1

Drageset, O. G. (2015). Student and teacher interventions: A framework for analysing mathematical discourse in the classroom. *Journal of Mathematics Teacher Education, 18*(3), 253–272. https://doi.org/10.1007/s10857-014-9280-9

Enders, C. K. (2010). *Applied missing data analysis.* Guildford.

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

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
Eriksson, E., Björklund Boistrup, L., & Thornberg, R. (2017). A categorisation of teacher feedback in the classroom: A field study on feedback based on routine classroom assessment in primary school. *Research Papers in Education, 32*(3), 316–332. https://doi.org/10.1080/02671522.2016.1225787
Fyfe, E. R., & Brown, S. A. (2018). Feedback influences children's reasoning about math equivalence: A meta-analytic review. *Thinking & Reasoning, 24*(2), 157–178. https://doi.org/10.1080/13546783.2017.1359208
Fyfe, E. R., Rittle-Johnson, B., & DeCaro, M. S. (2012). The effects of feedback during exploratory mathematics problem solving: Prior knowledge matters. *Journal of Educational Psychology, 104*(4), 1094–1108. https://doi.org/10.1037/a0028389
Ginsburg, H. P. (2009). The challenge of formative assessment in mathematics education: Children's minds, teachers' minds. *Human Development, 52*, 109–128.
Hanna, G. (2014). Mathematical proof, argumentation, and reasoning. In S. Lerman (Ed.), *Encyclopedia of mathematics education* (pp. 404–408). Springer Netherlands. https://doi.org/10.1007/978-94-007-4978-8_102.
Hargreaves, E., McCallum, B., & Gipps, C. (2000). Teacher feedback strategies in primary classrooms - new evidence. In A. Susan (Ed.), *Feedback for learning* (pp. 21–31). Routledge.
Hattie, J., & Clarke, S. (2019). *Visible learning feedback*. Routledge.
Hattie, J., & Gan, M. (2011). Instruction based on feedback. In *Handbook of research on learning and instruction* (pp. 263–285). Routledge.
Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112.
Hennessy, S., Howe, C., Mercer, N., & Vrikki, M. (2020). Coding classroom dialogue: Methodological considerations for researchers. *Learning, Culture and Social Interaction, 25*, Article 100404. https://doi.org/10.1016/j.lcsi.2020.100404
Hennessy, S., Rojas-Drummond, S., Higham, R., Márquez, A. M., Maine, F., Ríos, R. M., et al. (2016). Developing a coding scheme for analysing classroom dialogue across educational contexts. *Learning, Culture and Social Interaction, 9*, 16–44. https://doi.org/10.1016/j.lcsi.2015.12.001
Howe, C., & Abedin, M. (2013). Classroom dialogue: A systematic review across four decades of research. *Cambridge Journal of Education, 43*(3), 325–356.
Howe, C., Hennessy, S., Mercer, N., Vrikki, M., & Wheatley, L. (2019). Teacher–student dialogue during classroom teaching: Does it really impact on student outcomes? *The Journal of the Learning Sciences, 28*(4–5), 462–512.
Jain, S., Bruce, M. A., Stellern, J., & Srivastava, N. (2007). Self-efficacy as a function of attributional feedback. *Journal of School Counseling, 5*(4).
Jeannotte, D., & Kieran, C. (2017). A conceptual model of mathematical reasoning for school mathematics. *Educational Studies in Mathematics, 96*(1), 1–16.
Johnson, B., & Christensen, L. (2012). *Educational research: Quantitative, qualitative, and mixed approaches*. Sage.
Jones, S., Tanner, H., & Treadaway, M. (2000). Raising standards in mathematics through effective classroom practice. *Teaching Mathematics and Its Applications: An International Journal of the IMA, 19*(3), 125–134.
Knudsen, J., Lara-Meloy, T., Stevens Stallworth, H., & Wise Rutstein, D. (2014). Advice for mathematical argumentation. *Mathematics Teaching in the Middle School, 19*(8), 494–500. https://doi.org/10.5951/mathteacmiddscho.19.8.0494
Kramarski, B., & Mevarech, Z. R. (2003). Enhancing mathematical reasoning in the classroom: The effects of cooperative learning and metacognitive training. *American Educational Research Journal, 40*(1), 281–310.
Krauss, S., Bruckmaier, G., Lindl, A., Hilbert, S., Binder, K., Steib, N., et al. (2020). Competence as a continuum in the COACTIV study: The "cascade model". *ZDM, 52*(2), 311–327. https://doi.org/10.1007/s11858-020-01151-z
Lim, W., Lee, J.-E., Tyson, K., Kim, H.-J., & Kim, J. (2020). An integral part of facilitating mathematical discussions: Follow-up questioning. *International Journal of Science and Mathematics Education, 18*(2), 377–398. https://doi.org/10.1007/s10763-019-09966-3
Linacre, J. M. (2021). *Facets Rasch measurement computer program*. Retrieved 10.06.21 from Version 3.83.6. https://www.winsteps.com/facets.htm.
Lithner, J. (2000). Mathematical reasoning in task solving. *Educational Studies in Mathematics, 41*(2), 165–190. http://www.jstor.org/stable/3483188.
Lüken, M. M., Peter-Koop, A., & Kollhoff, S. (2014). *Influence of early repeating patterning ability on school mathematics learning joint meeting of the international group for the psychology of mathematics education (PME) (38th) and the north American chapter of the psychology of mathematics education (PME-NA)*. Vancouver, Canada, Jul 15-20, 2014.
Medway, F. J., & Venino, G. R. (1982). The effects of effort feedback and performance patterns on children's attributions and task persistence. *Contemporary Educational Psychology, 7*(1), 26–34. https://doi.org/10.1016/0361-476X(82)90004-2
Mercer, N., & Sams, C. (2006). Teaching children how to use language to solve maths problems. *Language and Education, 20*(6), 507–528. https://doi.org/10.2167/le678.0
Meusen-Beekman, K. D., Joosten-ten Brinke, D., & Boshuizen, H. P. A. (2016). Effects of formative assessments to develop self-regulation among sixth grade students: Results from a randomized controlled intervention. *Studies In Educational Evaluation, 51*, 126–136. https://doi.org/10.1016/j.stueduc.2016.10.008
Moschkovich, J. (2007). Examining mathematical discourse practices. *For the Learning of Mathematics, 27*(1), 24–30. http://www.jstor.org/stable/40248556.
Munson, J. (2019). After eliciting: Variation in elementary mathematics teachers' discursive pathways during collaborative problem solving. *The Journal of Mathematical Behavior, 56*, Article 100736. https://doi.org/10.1016/j.jmathb.2019.100736
Muthén, L. K., & Muthén, B. O. (2017). *Mplus user's guide* (8th ed.). Muthén & Muthén.
Panadero, E., & Jonsson, A. (2013). The use of scoring rubrics for formative assessment purposes revisited: A review. *Educational Research Review, 9*(0), 129–144. https://doi.org/10.1016/j.edurev.2013.01.002
Perry, N. E. (1998). Young children's self-regulated learning and contexts that support it. *Journal of Educational Psychology, 90*(4).
Pintrich, P. R., & De Groot, E. V. (1990). Motivational and self-regulated learning components of classroom academic performance. *Journal of Educational Psychology, 82*(1), 33.
Rakoczy, K., Pinger, P., Hochweber, J., Klieme, E., Schütze, B., & Besser, M. (2019). Formative assessment in mathematics: Mediated by feedback's perceived usefulness and students' self-efficacy. *Learning and Instruction, 60*, 154–165. https://doi.org/10.1016/j.learninstruc.2018.01.004
Ramaprasad, A. (1983). On the definition of feedback. *Behavioral Science, 28*(1).
Ruiz-Primo, M. A. (2007). *Exploring teachers' informal formative assessment practices and students' understanding in the context of scientific inquiry* (Vol. 44), 1.
Ruiz-Primo, M. A. (2011). Informal formative assessment: The role of instructional dialogues in assessing students' learning. *Studies In Educational Evaluation, 37*, 15–24.
Ruiz-Primo, M. A., & Brookhart, S. M. (2018). *Using feedback to improve learning*. Routledge.
Ruiz-Primo, M. A., & Furtak, E. M. (2007). Exploring teachers' informal formative assessment practices and students' understanding in the context of scientific inquiry. *Journal of Research in Science Teaching, 44*(1), 57–84.
Ruiz-Primo, M. A., & Min, L. (2013). Examining formative feedback in the classroom context: New research perspectives. In J. H. McMillan (Ed.), *SAGE Handbook of research on classroom assessment* (pp. 215–232). SAGE Publications, Inc. https://doi.org/10.4135/9781452218649.
Schoenfeld, A. H. (2015). Summative and formative assessments in mathematics supporting the goals of the common core standards. *Theory Into Practice, 54*(3), 183–194. https://doi.org/10.1080/00405841.2015.1044346
Schwabe, F., Trendtel, M., & McElvany, N. (2019). Die Bedeutung des sozioökonomischen Hintergrunds für spezifische Leseleistungsunterschiede bei offenen und geschlossenen Aufgaben und verschiedenen Textarten. *Zeitschrift für Grundschulforschung, 12*(1), 49–65. https://doi.org/10.1007/s42278-018-00036-1
Semadeni, Z. (1984). Action proofs in primary mathematics teaching and in teacher training. *For the Learning of Mathematics, 4*(1), 32–34. http://www.jstor.org/stable/40247842.
Sfard, A. (2001). There is more to discourse than meets the ears: Looking at thinking as communicating to learn more about mathematical learning. *Educational Studies in Mathematics, 46*(1/3), 13–57. http://www.jstor.org/stable/3483239.
Shilo, A., & Kramarski, B. (2019). Mathematical-metacognitive discourse: How can it be developed among teachers and their students? Empirical evidence from a videotaped lesson and two case studies [journal article]. *ZDM, 51*, 625–640. https://doi.org/10.1007/s11858-018-01016-6
Smit, R. (2009). *Die formative Beurteilung und ihr Nutzen für die Entwicklung von Lernkompetenz*. Schneider Verlag Hohengehren.
Smit, R., Bachmann, P., Blum, V., Birri, T., & Hess, K. (2017). Effects of a rubric for mathematical reasoning on teaching and learning in primary school. *Instructional Science, 45*(5), 603–622. https://doi.org/10.1007/s11251-017-9416-2
Smit, R., & Birri, T. (2014). Assuring the quality of standards-oriented classroom assessment with rubrics for complex competencies. *Studies In Educational Evaluation, 43*(December), 5–13. https://doi.org/10.1016/j.stueduc.2014.02.002
Smit, R., Dober, H., Hess, K., Bachmann, P., & Birri, T. (2022). Supporting primary students' mathematical reasoning practice: The effects of formative feedback and the mediating role of self-efficacy. *Research in Mathematics Education*, 1–24. https://doi.org/10.1080/14794802.2022.2062780
Squire, K. D., MaKinster, J. G., Barnett, M., Luehmann, A. L., & Barab, S. L. (2003). Designed curriculum and local culture: Acknowledging the primacy of classroom culture. *Science Education, 87*(4), 468–489.
Stovner, R. B., & Klette, K. (2022). Teacher feedback on procedural skills, conceptual understanding, and mathematical practices: A video study in lower secondary mathematics classrooms. *Teaching and Teacher Education, 110*, Article 103593. https://doi.org/10.1016/j.tate.2021.103593
Stylianides, G. J. (2008). An analytic framework of reasoning-and-proving. *For the Learning of Mathematics, 28*(1), 9–16.
Swiss Conference of Cantonal Ministers of Education (EDK). (2011). *Basic competences in mathematics. National standards*. Bern: EDK Retrieved from http://edudoc.ch/record/96784/files/grundkomp_math_d.pdf.
Tait-McCutcheon, S. L. (2008). *Self-efficacy in mathematics: Affective, cognitive, and conative domains of functioning, 1*, 507–513.
Tanner, H., & Jones, S. (2003). Self-efficacy in mathematics and students' use of self-regulated learning strategies during assessment events. *Psychology of Mathematics Education, 27* (Honolulu, Hawai'i, USA).
Van der Schaaf, M., Baartman, L., Prins, F., Oosterbaan, A., & Schaap, H. (2013). Feedback dialogues that stimulate students' reflective thinking. *Scandinavian Journal of Educational Research, 57*(3), 227–245. https://doi.org/10.1080/00313831.2011.628693
Verhoeven, L., & van Elsäcker, W. (2016). Home and school predictors of reading achievement in linguistically diverse learners in the intermediate primary grades. In *Written and spoken language development across the lifespan* (pp. 65–76). Springer.
Verschaffel, L., Schukajlow, S., Star, J., & Van Dooren, W. (2020). Word problems in mathematics education: A survey. *ZDM, 52*(1), 1–16. https://doi.org/10.1007/s11858-020-01130-4
Viholainen, A. (2011). The view of mathematics and argumentation, 9th-13th February. In *Proceedings of the 7th congress of the European society for research in mathematics education*. Poland: University of Rzeszów.

Von Davier, M., Gonzalez, E., & Mislevy, R. (2009). What are plausible values and why are they useful. *IERI monograph series, 2*, 9–36.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes.* Harvard University Press.

Webb, N. M., Franke, M. L., Ing, M., Turrou, A. C., Johnson, N. C., & Zimmerman, J. (2017). Teacher practices that promote productive dialogue and learning in mathematics classrooms. *International Journal of Educational Research.* https://doi.org/10.1016/j.ijer.2017.07.009

Wheatley, G. H. (1992). The role of reflection in mathematics learning. *Educational Studies in Mathematics, 23*(5), 529–541.

Whitenack, J., & Yackel, E. (2002). Making mathematical arguments in the primary grades: The importance of explaining and justifying ideas. *Teaching Children Mathematics, 8*(9), 524–528.

Wiliam, D., & Thompson, M. (2007). Integrating assessment with learning: What will it take to make it work? In C. A. Dwyer (Ed.), *The future of assessment: Shaping teaching and learning* (pp. 53–82). Lawrence Erlbaum Associates.

Winstone, N. E., Nash, R. A., Parker, M., & Rowntree, J. (2017). Supporting learners' agentic engagement with feedback: A systematic review and a taxonomy of recipience processes. *Educational Psychologist, 52*(1), 17–37. https://doi.org/10.1080/00461520.2016.1207538

Wright, B., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*(3), 370.

Wu, Y., & Schunn, C. D. (2021). From plans to actions: A process model for why feedback features influence feedback implementation. *Instructional Science, 49*(3), 365–394. https://doi.org/10.1007/s11251-021-09546-5

Wyatt-Smith, C., & Adie, L. (2019). The development of students' evaluative expertise: Enabling conditions for integrating criteria into pedagogic practice. *Journal of Curriculum Studies*, 1–21. https://doi.org/10.1080/00220272.2019.1624831

Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. *Theory Into Practice, 41*(2), 64–70.