Learning and Instruction 82 (2022) 101668

Contents lists available at ScienceDirect

Learning and Instruction

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

How do professional development programs on comparing solution methods and classroom discourse affect students' achievement in mathematics? The mediating role of students' subject matter justifications

Sog Yee Mok[a,*], Christian S. Hämmerle[a], Christian Rüede[b], Fritz C. Staub[a]

[a] University of Zurich, Switzerland
[b] University of Applied Sciences and Arts Northwestern Switzerland, Switzerland

ARTICLE INFO

Keywords:
Comparing solution methods
Classroom talk
Teachers' professional development
Reasoning
Mathematics achievement

ABSTRACT

Comparing solution methods fosters strategy flexibility in equation solving. Productive classroom discourse such as Accountable Talk (AT) orchestrated by teachers can improve students' justifications during classroom discussions and achievement. Do students' subject matter justifications during classroom discourse mediate the effect of teachers' professional development (PD) programs focused on comparing and AT on students' mathematics achievement? We investigated whether two PD programs (comparing or comparing+AT) compared to a control group increased the number of students justifications, and whether this affected mathematics achievement (strategy flexibility, procedural knowledge, and conceptual knowledge). The study (739 9th and 10th grade students in 39 classes) had an experimental pre-post control group design. Both PD programs significantly increased students justifications compared to the control group. The results of our multilevel path models showed significant small mediation effects in the comparing+AT group on procedural and conceptual knowledge. No mediation effects were found in the comparing group.

1. Introduction

Equation solving is a pivotal topic in mathematics classrooms (e.g., National Council of Teachers of Mathematics, 2006; Rüede & Staub, 2019). Using comparing in equation solving can support learning, as students elaborate on multiple correct solution methods to solve a problem (e.g., Rittle-Johnson & Star, 2007). Comparing solution methods means that students compare two correct solution methods (also known as solution strategies; Durkin, Star, & Rittle-Johnson, 2017) presented as worked-out examples. While comparing both methods students are asked to explain similarities and differences or to evaluate the efficiency across the different solution methods (Rittle-Johnson, Star, & Durkin, 2020). Many studies have shown that comparing solution methods increases students' strategy flexibility (Durkin et al., 2017; Verschaffel, Luwel, Torbeyns, & Van Dooren, 2009), which is defined as the ability to describe, evaluate, and enact multiple solutions in equations (Rittle-Johnson, Star, & Durkin, 2012; Verschaffel et al., 2009). Procedural and conceptual knowledge are also associated with strategy flexibility (Schneider, Rittle-Johnson, & Star, 2011). Laboratory studies focusing on comparing (Durkin et al., 2017) have found positive effects on these types of knowledge; however, field studies in classrooms are scarce (Durkin, Rittle-Johnson, Star, & Loehr, 2021; Rüede, Mok, & Staub, In press; Star et al., 2015).

Teachers' orchestration of classroom discussion can be productive for student learning and is known as productive classroom discourse (Walshaw & Anthony, 2008). From a sociocultural perspective, the cognitive development and knowledge of students will be promoted when students make relevant verbal contributions during classroom discourse and interact with others (Resnick, Asterhan, Clarke, & Schantz, 2018; Vygotsky, 1978). Verbal contributions such as justifications are expected to improve student learning especially in mathematics. Students justify a result or claim by presenting reasons to convince themselves or persuade others (Mata-Pereira & da Ponte, 2017; Sowder & Harel, 1998). Thus, with justifications students answer the question of why something works.

Different professional development (PD) programs have focused on promoting classroom discussions. Less is known about whether PD programs focusing on comparing solution methods and productive classroom discourse influence student justifications, and whether these in turn affect student achievement in mathematics. The present study

* Corresponding author. Kantonsschulstrasse 3, 8001, Zurich, Switzerland.
E-mail address: sogyee.mok@ife.uzh.ch (S.Y. Mok).

https://doi.org/10.1016/j.learninstruc.2022.101668
Received 15 July 2021; Received in revised form 1 July 2022; Accepted 9 August 2022
Available online 1 September 2022

addresses this research gap.

1.1. Comparing solution methods

1.1.1. Comparing solution methods and its effects on procedural knowledge, conceptual knowledge, and strategy flexibility in algebra

Equation solving can be conceptualized as problem solving (Schoenfeld, 1987). To develop equation solving successfully, the mathematics education literature has emphasized the importance of explicitly addressing *structures of equations* (e.g., Kieran, 1989; Linchevski & Livneh, 1999), *solution steps* (e.g., Cooper & Sweller, 1987), and *solution strategies* (e.g., Schoenfeld, 1987). In detail, discussing structures clarifies the problem (what is given?), discussing solution steps addresses the enactment and correctness of problem-solving steps, and discussing solution strategies supports planning, monitoring, and evaluating the problem-solving process (which strategy is efficient and why?). Knowledge about equation solving, i.e., about structures, solution steps, and strategies, can be conceptualized as a combination of procedural knowledge, conceptual knowledge, and strategy flexibility (Durkin et al., 2021; Schneider et al., 2011). These three knowledge types are related but can be measured independently of each other (Schneider et al., 2011). Procedural knowledge is defined as the ability to perform sequences of actions to solve problems (Rittle-Johnson & Star, 2009). Conceptual knowledge is the ability to recognize and explain central concepts of a domain (Rittle-Johnson & Star, 2009). *Strategy flexibility* refers to knowledge about multiple solution methods and how to compare them with respect to their suitability (Rittle-Johnson & Star, 2009; Verschaffel et al., 2009). *Strategy flexibility* can be further distinguished into flexibility knowledge and flexibility use. Flexibility *knowledge* is defined as knowledge about multiple solution methods and their suitability for efficiently solving equations. It enables students to describe a broad range of applications for different solution methods and to give reasons for or against the suitability of a chosen method. Flexibility *use* is defined as the ability to select and implement the most appropriate solution method among different methods (Rittle-Johnson et al., 2012). Research suggests that a high level of strategy flexibility is associated with higher levels of procedural and conceptual knowledge (Schneider et al., 2011).

Middle and high school students often struggle with equation solving (Booth, Eyer, & Paré-Blagoev, 2014). *Comparing* solution methods can help students to learn to solve equations efficiently. In this approach, students learn to compare multiple correct solution methods as worked-out examples by assessing which method is most efficient for solving a problem and providing reasons for their choice (Rittle-Johnson & Star, 2007). Different laboratory studies using comparing solution methods have shown how students' strategy flexibility, procedural knowledge, and conceptual knowledge in algebra can be improved (Rittle-Johnson & Star, 2007; Rittle-Johnson & Star, 2009). To date, only three PD studies have focused on comparing solution methods, and they have returned mixed findings regarding students' achievement in algebra (Durkin et al., 2021; Rüede et al., in press; Star et al., 2015). In a study by Star et al. (2015), teachers in a PD were trained to use comparing tasks and were provided with supplementary material for their algebra lessons. The study showed no effect of the PD on students' achievement outcomes (Star et al., 2015). A recent study by Durkin et al. (2021) developed a PD and supplementary instructional material for teachers to support students' discussions and math achievement. The instructional materials included worked example pairs including question prompts (e.g., "which is better?"). In the PD, teachers were introduced to the materials, viewed videotaped examples of lessons, and had the opportunity to plan, teach, and discuss sample lessons with peer teachers. The overall achievement measure consisted of three subscales: procedural knowledge, conceptual knowledge, and procedural flexibility (which we call strategy flexibility). The results showed higher scores for students in the PD group on the overall achievement measure compared to students in the control group. The advantage for students in the PD group was mainly due to higher procedural flexibility. Another study[1] found that students with teachers who participated in one of two PD programs focused on comparing or comparing plus productive classroom discourse showed significantly higher achievement gains on flexibility use, flexibility knowledge, and procedural knowledge from pre- to posttest than students in the control group. Students in the PD group focused on comparing plus productive classroom discourse also improved their conceptual knowledge (Rüede et al., in press).

1.1.2. Comparing solution methods and justifications

The use of tasks in classroom discussions that demand cognitive activation are essential for fostering high-level thinking and reasoning (Resnick, Michaels, & O'Connor, 2010; Stein, Grover, & Henningsen, 1996). Comparing solution methods tasks have the potential to cognitively activate students and facilitate their learning and reasoning. For instance, comparing tasks can offer different prompts to support reflection and discussions about differences and similarities between the solution methods (Durkin et al., 2021; Rittle-Johnson & Star, 2007). Some of the prompts present why-questions. Why-questions can lead students in mathematics classroom discussions to justify, for example, why one solution works better than another (Kazemi & Stipek, 2009). Even though comparing solution methods offer opportunities to provide justifications, very little is known about the effects of comparing solution methods on student justifications during classroom discourse. Nevertheless, based on theoretical assumptions (Resnick et al., 2010; Stein et al., 1996) and the design of comparing tasks with prompts (e.g., Rittle-Johnson & Star, 2007, 2009), we assume that the use of comparing solution methods tasks can positively affect student justifications.

1.2. Accountable Talk and its influence on student achievement and justifications

From a sociocultural perspective, the development of knowledge in classroom discourse can be promoted when students make relevant verbal contributions during classroom discourse and build on each other's ideas (Resnick et al., 2018). Based on this perspective, classroom discourse is productive for student learning when the classroom discourse fosters student thinking (Kim & Wilkinson, 2019; Walshaw & Anthony, 2008). One well-established approach to promoting productive classroom discourse is called *Accountable Talk* (AT; Michaels, O'Connor, & Resnick, 2008). According to AT, three types of accountabilities provide a framework for organizing productive classroom discourse (Michaels et al., 2008, 2010): accountability to the learning community, accountability to accurate knowledge, and accountability to rigorous thinking. Students are accountable to the learning community if they listen to each other and build their utterances on one another's ideas during discourse. Students are accountable to accurate knowledge when classmates try to be accurate and appropriate in their utterances. To achieve this accountability, students might offer an explanation or example that is grounded in knowledge or facts. Students are accountable to rigorous thinking when they build a line of argumentation. To build compelling arguments, students must link claims and evidence in a coherent, logical, and rigorous way. A set of teacher moves (i.e., questions to challenge students' answers and press for accuracy) are offered to stimulate students' thinking and reasoning in classroom discourse in ways heeding these accountabilities (Michaels et al., 2008).

[1] This study by Rüede et al. (in press) is embedded in the same project of the present study. Rüede et al. (in press) focused on the professional development programs' effects on students' achievement in the posttest and follow-up-test.

1.2.1. PD effects of productive classroom discourse approaches on students' achievement

AT, among other approaches intended to foster productive classroom discourse, has been taught in many PD programs to improve classroom discourse outcomes and/or student achievement. Multiple PD studies found that the implementation of AT led to an increase in student achievement (e.g., Chen, Chan, Chan, Clarke, & Resnick, 2020; O'Connor, Michaels, Chapin, & Harbaugh, 2017; Rüede et al., in press). For instance, recent research showed positive effects of teachers' use of AT moves in classrooms on student achievement (Chen et al., 2020; O'Connor et al., 2017). In contrast, other studies using similar classroom talk approaches did not find positive achievement effects for students after introducing to productive classroom discourse practices (Pehmer, Gröschner, & Seidel, 2015; Reznitskaya et al., 2012; van der Veen, de Mey, van Kruistum, & van Oers, 2017).

1.2.2. PD effects of productive classroom discourse approaches on students' classroom discourse outcomes

Mixed findings have also been reported for the effects of PD focusing on improving classroom discourse with regard to students' classroom discourse outcomes. The majority of PD studies focusing on productive classroom discourse found positive changes in students' classrooms dialogue patterns (Chen et al., 2020). Sedova, Sedlacek, and Svaricek (2016) revealed a significant increase in students' reasoning after teachers attended a PD on how to teach productive classroom discourse. Reasoning in this study was defined as a complete thought including an explanation of the students' thinking. Other studies did not reveal improvements in classroom discourse (Pehmer et al., 2015; Pimentel & McNeill, 2013).

1.3. Role of justifications for mathematics achievement

Different definitions and labels in classroom discourse research have been used to describe students' verbal utterances related to reasoning, argumentation, explanations, and justifications (Asterhan & Schwarz, 2016; Mata-Pereira & da Ponte, 2017; Webb et al., 2014). In research focusing on building mathematical arguments during discourse, two subtypes can be highlighted (Asterhan & Schwarz, 2016; Kazemi & Stipek, 2009): explanations and justifications. Explanations have a clarifying function within discourse (Asterhan & Schwarz, 2016) and are used to describe solution steps to solve a problem. For example, students' explanations of procedural steps can be given without expressing why these steps work mathematically (Kazemi & Stipek, 2009, p. 126). Justifications, on the other hand, answer the question of why mathematical procedures work: Students justify a result or claim by presenting reasons to convince themselves or persuade others (Mata-Pereira & da Ponte, 2017; Sowder & Harel, 1998). Classroom discourse in mathematics that meets the standards of Accountable Talk must include subject-specific rigorous thinking in sharing and discussing justifications of problem solutions (Michaels et al., 2008). Students' justifications in mathematics need to be based on accepted concepts, procedures, properties, mathematical ideas, and arguments (Kazemi & Stipek, 2009; Mata-Pereira & da Ponte, 2017). Justifications are required when comparing solutions methods in order to discuss similarities and differences between solution steps across different solutions methods. For instance, to answer the prompts "which method is better? And why is it better?" (e.g., Durkin et al., 2021), students can provide justifications with regard to the efficiency of a solution method and relations between procedures or concepts (Mata-Pereira & da Ponte, 2017).

PD for teachers can influence different student outcomes during mathematics lessons such as student justifications, student explanations and generation of strategies. For instance, the level of completeness can also play a role when exploring explanations and justifications. Various studies differentiated between complete and incomplete explanations (e.g., Franke et al., 2009; Webb et al., 2008). However, this differs from the role completeness plays in justifications during classroom discourse (Inglis & Mejia-Ramos, 2008). According to AT, complete and incomplete justifications both provide opportunities for students to advance rigorous mathematical thinking in classroom discourse settings by allowing students to collaboratively build justifications (Michaels et al., 2008). In detail, following the accountability to rigorous thinking in AT, students can learn from the incomplete ideas expressed in justifications if other students or teachers detect why these are mathematically inadmissible or incomplete, and argue for a correct or more rigorous justification. Thus, we assume that justifications can achieve more depth in learning by building on others' (incomplete) ideas or refuting an idea. Supporting this, a study by Webb et al. (2014) found that the highest level of students' engagement (i.e., defined as adding to another student's proposed approach by suggesting further details to that approach or by suggesting an alternative approach), predicted achievement over and above providing explanations (Webb et al., 2014).

To sum up, we argue that the completeness of a justification is less important than students having the learning opportunity to put forward reasons against or for other students' justifications (or other utterances) to promote learning in classroom discourse. Thus, both incomplete and complete justifications can be beneficial for learning. Moreover, we concentrated on all mathematical justifications in the present work because we assume that they can affect especially students' conceptual knowledge (Kazemi & Stipek, 2009), which can also influence procedural knowledge and strategy flexibility (Schneider et al., 2011).

1.4. Subject matter justifications as mediator of PD effects on students' mathematics achievement

Effective PD programs need to increase teachers' knowledge and change teachers' instructional practices before these practices can improve student learning in the classroom as a long-term goal (Desimone, 2009). Following Desimone's (2009) model and sociocultural theories regarding the influence of productive classroom discourse on student justifications and learning (Michaels et al., 2008), we assume that effective PD programs have an effect on student justifications during classroom talk, which in turn affect student learning and achievement. Even though student learning should be the goal of each PD and changes in teaching are likely to affect student learning (Kennedy, 2016), only few studies have investigated the mediational role of student justifications on student achievement after teachers attended a PD (e.g., Sedova et al., 2019). Many studies have explored the effect of PD programs on students' reasoning and achievement separately (Reznitskaya et al., 2012; van der Veen et al., 2017). Some did not find improvements in achievement (Reznitskaya et al., 2012). A possible explanation for the lack of improvement in achievement may be related to the design of the PD programs focused on productive classroom discourse in general, as they focused on a broader learning content that could apply to teachers in the PD who were specialized in different school subjects (e.g., Pehmer et al., 2015).

In the present study, we assume that PD programs for comparing solution methods will be most effective for students' learning processes when combined with instruction on orchestrating productive classroom discourse based on AT. Through the use of comparing solution methods in solving equations, students are provided with learning opportunities by answering prompts to compare different solutions of the same problem and to give justifications for their selection of efficient methods (Durkin et al., 2017; Rittle-Johnson & Star, 2007). Thus, we assume that comparing solution methods with prompts can foster justifications, but additionally supporting comparing with AT during classroom discourse should facilitate the generation of justifications even more.

Some research studying comparing in real classrooms without providing explicit training on orchestrating productive classroom discourse failed to reveal positive effects on strategy flexibility or procedural and conceptual knowledge (e.g., Star et al., 2015). Providing teachers with opportunities to learn how to orchestrate classroom discourse based on AT can promote rigorous thinking and subject matter learning. For instance, with AT moves, teachers can initiate and scaffold discussions of students' subject matter justifications regarding important terms and mathematical arguments in equation solving (e.g., rules) while comparing two solution methods as worked-out examples. Subject matter justifications in equation solving refer to justifications regarding recognizing structures, enacting solutions steps, and planning and evaluating solution strategies to solve equations (see 1.1). Students can build on subject matter justifications in mathematical arguments by elaborating and relating ideas of their own to mathematically inadmissible or incomplete justifications (Webb et al., 2014; see 1.3). Teachers can use AT moves to press for students to answer the question of why solution methods work when comparing multiple solution methods (cf. Kazemi & Stipek, 2009; Litke, 2020). Linking related procedures and mathematical concepts is essential to justifying why solution methods work (Mata-Pereira & da Ponte, 2017).

Studies about the mediation of the effects of classroom discourse PD programs on student achievement by students' subject matter justifications has been scarce. Nevertheless, these studies support our assumption that the orchestration of productive classroom discourse based on AT when comparing can be beneficial for supporting students' sharing and discussion of subject matter justifications in mathematics, which we expect to positively affect achievement. Matsumura, Garnier, and Spybrook (2013) showed a positive effect of their PD on students' literacy achievement which was mediated by the quality of classroom text discussions based on AT principles. The classroom text discussions were guided by teachers who used the discussion method "Questioning the Author" (Beck & McKeown, 2006). Teachers in this study developed open questions regarding the literacy texts to support students' text comprehension and to let students make connections between the content and relevant concepts (Matsumura et al., 2013). A study conducted in regular mathematics lessons found similar results, showing that the relationship between talk-supportive teacher practices and students' mathematics learning was mediated by students' verbal participation during classroom discourse (Ing et al., 2015). Taken together, both studies (Ing et al., 2015; Matsumura et al., 2013) revealed that teachers' instructional practices helped improve students' classroom discourse quality by making connections between content and concepts.

We hypothesize a mediation effect for teachers using comparing solution methods (even without the AT component) because the comparing tasks offered prompts that can support justifications. Drawing on the mediation findings mentioned and the assumptions of sociocultural theories (Michaels et al., 2008), we furthermore assume that the combination of teachers' use of comparing solution methods and the orchestration of productive talk will lead to learning opportunities for students' subject matter justifications, which in turn will affect student achievement (i.e., mediation effect). However, we expect the mediation effect for the intervention that combined comparing solution methods with AT to be larger than for the comparing group only. Moreover, as conceptual knowledge is associated with procedural knowledge and strategy flexibility (Schneider et al., 2011), we assume that justifications will mediate the effect of PD on all three knowledge types.

1.5. Aims of the study and research questions

The present study seeks to contribute to the literature on effective PD programs (Desimone, 2009) by investigating students' justifications during classroom discourse as a mediator of the effects of PD programs focused on comparing solution methods and AT on students' achievement in algebra (i.e., strategy flexibility, procedural knowledge, and conceptual knowledge). We investigate two sets of research questions (RQs) regarding the PD effects on students justifications and indirect effects of PD programs via justifications on achievement: RQ1a) Do students in the PD comparing group generate more justifications during classroom discourse than students in the control group (CG)? RQ1b) Do students in the PD comparing and AT (i.e., comparing+AT) group generate more justifications during classroom discourse than students in the CG? RQ1c) Do students in the PD comparing+AT group generate more justifications during classroom discourse than students in the PD comparing group? RQ2a) Do students justifications mediate the effect of the PD comparing group on students' strategy flexibility, procedural knowledge, and conceptual knowledge in comparison to the CG? RQ2b) Do students justifications in classrooms mediate the effect of the PD comparing+AT on students' strategy flexibility, procedural knowledge, and conceptual knowledge in comparison to the CG? RQ2c) Do the mediation effects of the PD comparing+AT group exceed the mediation effects of the PD comparing group?

We hypothesize that students in both PD groups (comparing and comparing+AT) will generate significantly more justifications than students in the CG (PD comparing: *Hypothesis 1a*; PD comparing+AT: *Hypothesis1b*). Moreover, we hypothesize that students in the PD comparing+AT group will generate significantly more justifications than those in the PD comparing group (*Hypothesis 1c*). We also hypothesize that for students in both PD groups (compared to students in the CG) students' justifications will significantly mediate the effect of the PD on all student achievement outcomes (PD comparing: *Hypothesis 2a*; comparing+AT: *Hypothesis 2b*). Finally, we hypothesize that the mediation effects in the PD comparing+AT group will be significantly larger than in the PD comparing group (*Hypothesis 2c*).

2. Method

2.1. Design and participants

We used an experimental design with two experimental groups (EGs, i.e., a PD either on comparing alone (EG_{comp}) or comparing+AT ($EG_{comp+AT}$)) and a wait-list CG. Interested mathematics teachers registered on a voluntary basis for a PD program on quadratic equation solving. Teachers were randomly assigned to EG_{comp}, $EG_{comp+AT,}$ or CG. Students' flexibility knowledge scores were assessed at different time points. We measured students' prior knowledge with a baseline test (t_0) before the PD programs, in the first three weeks of the school year. To give teachers the opportunity to practice planning lessons and to try the content of the PD programs, they taught two lessons in their math classes that was based on the concept learned but did not focus on quadratic equation solving. We then measured students' knowledge with a pretest (t_1) at the beginning of the teaching unit about quadratic equation solving and a posttest (t_2) at the end of the teaching unit. The teaching unit included 16 lessons. The researchers of this study randomly selected four lessons out of the 16 lessons in the unit on quadratic equations which were videotaped (for an overview of the PD program and achievement measures, see Fig. 1). 39 teachers (38.5% female; 13 teachers per group) and their classes participated in the study. All students participated voluntarily. Both students and their parents signed a consent form to be enrolled in the study. The 9th or 10th grade students attended either an upper secondary school or an upper vocational school in German-speaking cantons of Switzerland. The final sample consisted of $N = 739$ students (59% female, $M_{Age} = 15.8$ years, $SD_{Age} = 1.10$ years) in 39 classes.

2.2. Procedure of the PD programs

Teachers in the EG_{comp} attended a PD which aimed to improve students' flexible use of different solution methods by the use of comparing multiple solution methods as worked-out examples. Teachers in the $EG_{comp+AT}$ attended a PD that assisted teachers in focusing on subject matter content while comparing solution methods and on orchestrating productive classroom discourse (e.g., discussing which teacher moves help students to achieve the lesson goals; Michaels et al., 2008). Both PD programs supported teachers in focusing on subject matter content (i.e., recognizing structures, enacting solution steps, choosing strategies) while discussing similarities and differences between multiple solution methods (see Rüede et al., in press, for details). Discussing similarities

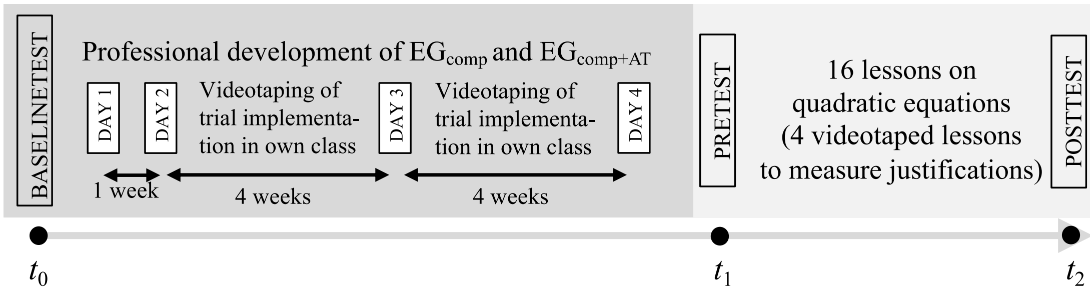

Fig. 1. Timeline of PD programs and measurement points for achievement tests.

and differences between different solutions of an equation (guided by teachers' focus on relevant subject matter content) provides students with learning opportunities to discuss justifications for selected solution methods. For instance, recognizing structures of equations (e.g., identifying the same compound term on both sides of an equation) is subject matter content that helps students discuss similarities in procedures in a productive way to enhance learning.

2.3. Measures

2.3.1. Intervention fidelity check variables

We explored two variables to evaluate the fidelity to the PD programs. The number of student turns related to solving quadratic equations during classroom discourse were coded based on the videotaped lessons. A higher number of student turns is assumed to be a distal indicator for productive classroom discourse. All tasks related to the solution of quadratic equations were coded for the occurrence of given worked-out solutions to be compared. A comparing task was coded with 1 if the solution methods were presented side-by-side. A frequent use of comparing tasks is considered an indicator for successful implementation of the PD comparing.

2.3.2. Main variables

In the following, we introduce the main variables used in our multilevel path models: EGs, students' justifications, and four achievement outcomes (flexibility use, flexibility knowledge, procedural knowledge, and conceptual knowledge).

2.3.2.1. Experimental groups. The two EGs (EG_{comp}; $EG_{comp+AT}$) were used as independent variables in our models. Both EGs were compared to CG. Thus, both groups were dummy-coded in our path models and the CG was the reference group.

2.3.2.2. Students' justifications. In the present study, we focused on justifications in classroom discourse with regard to the subject matter content, because explicit mathematical justifications are related to conceptual and procedural understanding (Legesse et al., 2020; Mata--Pereira & da Ponte, 2017). Students' justifications were used as a mediator variable in our models. Sequences of classroom discourse (within the four videotaped lessons) were identified and transcribed. We coded justifications generated by students across four videotaped lessons in a unit on quadratic equations.

A student's turn was coded as a justification when a student expressed at least one justification within equation solving.[2] To define subject matter justifications within equation solving, we connected justifications to the content regarding equation solving discussed in the PD programs (i.e., recognizing structures, enacting solution steps, or planning and evaluation of a (solution) strategy or solution method; (Rüede et al., in press). A transcript illustrating the codes used for justifications regarding recognizing structures, enacting solution steps, or planning and evaluation of a (solution) strategy is available in Table 1 (for more details about the coding of the subject matter justifications and further examples, see *Appendix A1*).

We computed the number of students' justifications as the sum of all students' turns with justifications related to the subject matter content in equation solving. 20% of the data was double-coded. The interrater

Table 1
Transcript of a classroom discourse example with justifications.

Is this a quadratic equation?
The equation $N^2+2N = 15$ is on the black board.
During the discussion the teacher adds notes to the black board.

#	Speaker	Turns	Code
1	T	Is this a quadratic equation?	
2	Lou	Yes.	
3	Sam	The solution is 3, **because** if you fill in 3 for *N* it is correct.	J-E
4	T	Ok. Everyone. The question was: Is this a quadratic equation?	
5	Kim	**As** *N* is squared, this is a quadratic equation.	J-S
6	T	Loraine.	
7	Loraine	Oftentimes you bring it to the form "*a x to the power of 2 plus b x plus c is equal to 0*" =	
8	T	= Be quiet Kim. Pardon me, Loraine, please continue.	
9	Loraine	= **To get** that form here, you should subtract 15 **that** it equals 0.	J-P
10	T	You would subtract 15 and you would have a quadratic equation. Questions?	
11	Alex	**Why** does it have to be equal to 0?	J-S
12	T	Loraine?	
13	Loraine	Why? **Because** afterwards you can factor with brackets =	J-E
14	Alex	= Ok.	
15	Loraine	= **If** it is factored and equals 0, you have *N* plus 5 and *N* minus 3 and then you see that the solutions are −5 and 3.	J-P
16	Alex	Ok.	
17	T	The question was if this is a quadratic equation. Now we have two equations. Are both equations quadratic equations or just one of them?	
18	Rafael	Both, so, both equations are the same.	
19	T	It is the same equation. Loraine, can you explain your thinking?	
20	Loraine	You try to bring a quadratic equation to this form. Afterwards you can factor.	
21	T	We just transformed the equation. Laura already thinks of solving the equation through factoring. We might be able to factor. The definition of a quadratic equation uses the *a*, *b* and *c*.	

Note. T = teacher; J = the code for students' justification; S stands for a turn recognizing a structure, E stands for a turn enacting a solution step, and P stands for a turn on planning and evaluation of a solution step or solution method. Reasoning words indicating justifications are indicated in bold; " = " indicates a turn when a speaker is getting interrupted or is interrupting another speaker.

[2] We additionally explored the number of incorrect justifications with a subsample (student turns with justifications in 9 randomly-selected classes, three from each group). 161 of the 802 turns with justifications were coded on whether the previously-coded justifications were mathematically correct or incorrect. Only one turn was classified as incorrect. Due to this very small number, we omitted further analyses on incorrect justifications.

reliability for students' justifications was substantial ($\kappa = 0.63$; Landis & Koch, 1977). We did not compute sums separately for justifications organized by content for the main analyses (i.e., mediation effects) because based on sociocultural theories and AT we assume that all justifications relevant for solving equations can contribute to student learning in equation solving and we are primarily interested in explaining mediation effects from an AT perspective.

2.3.2.3. Students' achievement scales after the teaching unit. Our achievement subscales at posttest were used as dependent variables in our models. Sample items for assessing procedural knowledge, flexibility use, flexibility knowledge, and conceptual knowledge in pretest and posttest were adapted from Rittle-Johnson and Star (2007; 2009). All sample items and the scoring of items are reported in *Appendix A2*.

We measured students' achievement after the 16 lessons within the quadratic equation unit were completed (in a posttest, t_2). We used the same scales and items as measured at t_1. The reliability of three of the subscales were good to very good, and the reliability of the conceptual knowledge subscale was acceptable (procedural knowledge t_2: 10 items; $\alpha = 0.74$; flexibility use t_2: 10 items; $\alpha = 0.87$; flexibility knowledge t_2: 28 items; $\alpha = 0.82$; conceptual knowledge t_2: 20 items; $\alpha = 0.56$).

2.3.3. Covariates

All baseline and pre-knowledge measures of our four achievement subscales served as covariates in our path models.

2.3.3.1. Baseline pre-knowledge at beginning of the school year. To consider potential baseline pre-knowledge of students with regard to the procedural knowledge, flexibility knowledge, flexibility use, and conceptual knowledge, we assessed these four types of baseline pre-knowledge at the beginning of the school year (in a baseline test, t_0) before teachers started the professional development workshops. This flexibility knowledge in the baseline test served as covariate in our analyses. The reliability of the subscales ranged from acceptable to good (procedural knowledge t_0: 7 items; $\alpha = 0.41$; flexibility use t_0: 7 items; $\alpha = 0.76$; flexibility knowledge t_0: 18 items; $\alpha = 0.71$; conceptual knowledge t_0: 14 items; $\alpha = 0.61$).

2.3.3.2. Pre-knowledge before the teaching unit. During the PD, teachers were encouraged to apply comparing solution methods in two different lessons on topics other than quadratic equations as practice lessons. Even though these practice lessons did not involve quadratic equations, it would be possible that teachers already adapted their teaching style to the content of the PD, leading to potential differences between the conditions. Thus, we also measured the four types of pre-knowledge directly before the start of the relevant unit of quadratic equations (in a pretest, t_1). The reliability of three of the subscales were very good to good, and the reliability of the conceptual knowledge subscale was acceptable (procedural knowledge t_1: 10 items; $\alpha = 0.71$; flexibility use t_1: 10 items; $\alpha = 0.85$; flexibility knowledge t_1: 28 items; $\alpha = 0.82$; conceptual knowledge t_1: 20 items; $\alpha = 0.52$).

2.4. Data analyses

Across all time points and dependent variables, the missing values in our data ranged from 4.7 to 7.4%. We used Full Information Maximum Likelihood procedure (FIML) in *Mplus* to impute the missing data (Enders, 2010). Multilevel path models were used because student data (Level 1) were nested in classes (Level 2).

A multilevel path model was tested for each outcome (for the hypothesized statistical model including all paths, see Fig. 2). In a mediation model the *a*-path indicates the relationship between the independent variable and the mediator variable, and the *b*-path indicates the relationship between the mediator variable and the dependent variable controlling for the independent variable (Preacher & Kelley, 2011). The *c'*-path indicates the relationship between the independent variable and the dependent variable after controlling for the mediator variable. The *c'*-path is called the "direct effect" (Fritz & MacKinnon, 2007). The product between the *a*-path and the *b*-path forms the "indirect effect" (Preacher & Kelley, 2011), which is considered the mediation effect. The relationship between the independent variable and the dependent variable is called the "total effect" (Fritz & MacKinnon, 2007), which can also be calculated as sum of the direct and indirect effect. We used the Robust Maximum Likelihood estimator. For each path analysis, we applied a "manifest measurement – latent

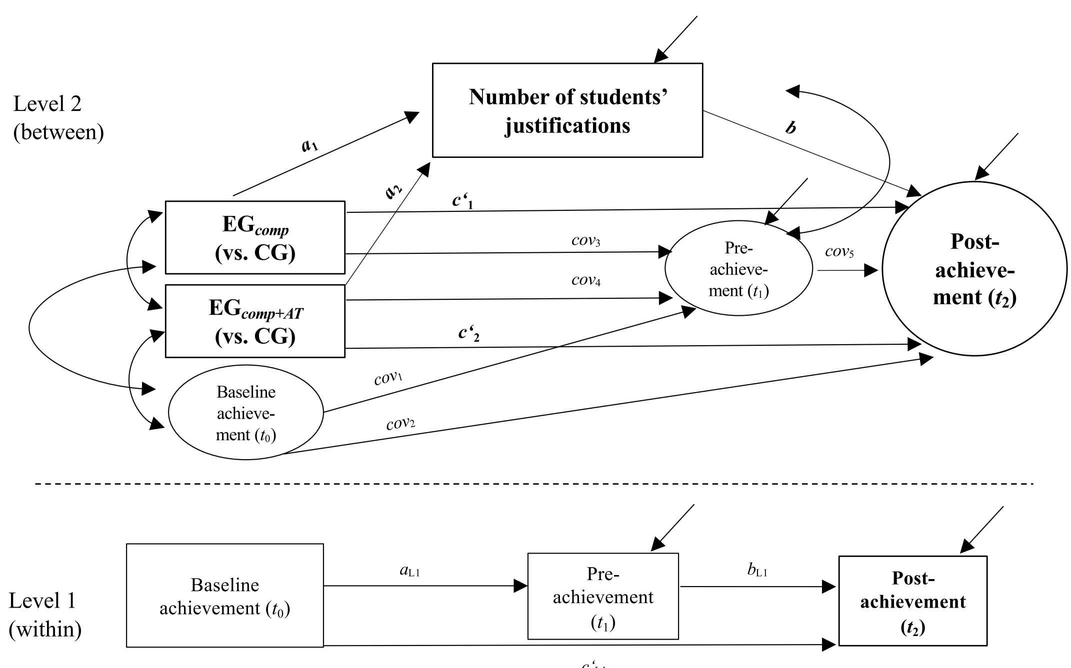

Fig. 2. Hypothesized multilevel path model using the "manifest measurement – latent aggregation" model approach by Lüdtke et al. (2011).
Notes. Our path analyses are specified for the within and between parts of the analysis. Direct relationships between variables are indicated with an arrow. If variables are assumed to be correlated, their correlations are presented with a bidirectional arrow. Residuals are displayed with short arrows.

Table 2
Group differences in the number of student turns and comparing tasks.

	Number of student turns related to comparing equation solving	Comparing tasks with worked-out examples
CG	118 (73)	0.08 (0.28)
EG_{comp}	211 (84)	4.85 (3.31)[a] **
$EG_{comp+AT}$	285 (123)[a] ***	3.62 (4.11)[a] *

Notes. *$p < .05$, **$p < .01$, ***$p < .001$; Numbers in parentheses are the standard deviations. [a] indicates a significant difference to the control group of the post hoc analysis. Number of student turns were based on 16 lessons. Comparing tasks were related to the four videotaped lessons.

aggregation" model approach (Lüdtke, Marsh, Robitzsch, & Trautwein, 2011), using measured variables (i.e., students' achievement scores) at the individual level and observed variables at the classroom level (i.e., EGs; students' justifications during observed classroom talk), and the implicit latent function in *Mplus* to compute latent aggregated measures of the achievement scores at the classroom level. We further controlled for students' baseline- and pretest measures. In the path models, we computed the effects from both EGs to the mediator (a_1-path and a_2-path) as well as the effects from the mediator to each of the four outcomes controlling for the independent variable (*b*-paths). We computed indirect effects for each EG, which is the product of the *a*-path and the *b*-path (EG_{comp}: a_1-path x *b*-path and $EG_{comp+AT}$: a_2-path x *b*-path) to test the predicted significance of the mediation effects for all four outcomes (Fritz & MacKinnon, 2007). Then, we computed the direct effects (EG_{comp}: c'_1-path and $EG_{comp+AT}$: c'_2-path). We also tested if the two indirect effects of the EGs differ from each other in *Mplus* by calculating a difference parameter between the indirect effects of our models and testing whether this new difference parameter differed from zero.

Finally, we calculated Cohen's *d* for the PD effects on justifications and a standardized mean difference (*SMD*) score for each PD group (vs. CG) for each indirect effect. For the *SMD*, the unstandardized scores of each outcome are standardized by the standard deviation of each outcome (Cheung, 2018, p. 4). Using *SMD* as an effect size is particularly recommended for indirect effects in complex studies with multiple groups and outcomes (Cheung, 2009, 2018). The common Cohen's *d* benchmarks denote small ($d = 0.2$), medium ($d = 0.5$), and large ($d = 0.8$) effects. The *SMD* scores in our study can be interpreted with the effect size benchmarks by Cohen (1988) adapted to account for the multiplication of path *a* and path *b* to calculate the indirect effects (i.e., small effect: $SMD = 0.14$, medium effect: $SMD = 0.36$, and large effect: $SMD = 0.51$; Cheung, 2009).

3. Results

3.1. Intervention fidelity check

We analyzed the number of student turns in the classroom discussions as an indicator of productive talk in terms of students' participation. As expected, students in the $EG_{comp+AT}$ showed significantly more turns than students in the CG. There was no significant difference in the number of turns between the EG_{comp} and the CG. The EG_{comp} and the $EG_{comp+AT}$ both completed a higher number of comparing tasks with worked-out examples than the CG (Table 2). Thus, our intervention content was successfully implemented in classrooms.

3.2. PD program effects on the occurrence of justifications with different subject matter content

Justifications varied between 1 and 43 justifications per classroom ($M = 18.42$, $SD = 10.64$) across groups (EGs and CG). As knowledge about recognizing structures, enacting solution steps, and planning and evaluation of solutions are important in solving equations (Schoenfeld, 1987), justifications can be linked to these three content types. We present the means and standard deviations of these justifications with different subject matter content separated by groups (Table 3).

The ANOVA results showed significant differences between groups for the justifications related to enacting solution methods, $F(2, 36) = 6.075$, $p = .005$, and justifications related to planning and evaluation, $F(2, 36) = 7.644$, $p = .002$. The posthoc test showed that more justifications related to enacting solution methods in the $EG_{comp+AT}$ ($M = 13.31$, $SD = 9.45$) compared to the CG ($M = 4.54$, $SD = 3.02$, $p = .016$). In the EG_{comp} ($M = 9.31$, $SD = 5.04$) more justifications related to enacting solution methods were provided compared to the CG ($p = .022$). Regarding justifications related to planning and evaluation, the posthoc test showed more justifications regarding planning and evaluation in the $EG_{comp+AT}$ ($M = 6.77$, $SD = 2.86$) compared to the CG ($M = 3.08$, $SD = 3.07$, $p = .011$). A similar pattern was found for justifications related to planning and evaluation in the EG_{comp} ($M = 7.39$, $SD = 3.18$) compared the CG ($p = .003$). No significant group differences were found for justifications related to recognizing structures (all $ps > .05$, see Table 3). When students justifications were summed, both EGs differed significantly from the CG ($F(2, 36) = 6.886$, $p = .003$; EG_{comp} vs. CG: $p = .024$, $d = 1.114$, $EG_{comp+AT}$ vs. CG: $p = .010$, $d = 1.286$; Table 3). The effect sizes of both of the EGs on students justifications were large. The EGs were not significantly different from each other in the individual justification types and the sum of justifications (Table 3).

3.3. Path models for mediation effects of justifications

The ICCs were between 0.15 and 0.41 for all achievement measures. Our multilevel path models reached overall good values of fit indices (> .95 for a good Comparative Fit Index (CFI); < .06 for a good Root Mean Square Error of Approximation (RMSEA; Hu & Bentler, 1999; for details, see *Appendix A3*). All correlations are displayed in *Appendix A4*. The means and standard deviations of all manifest measures per group are displayed in *Appendix A5*.

In our path models, we found significant positive effects of the PD on the number of students' justifications in both EG in comparison to the CG (EG_{comp}: $B = 7.846$, $SE = 2.653$, $p = .003$; $EG_{comp+AT}$: $B = 13.692$, *SE*

Table 3
Means, standard deviations, and effect sizes of all students justifications by content in equation solving.

	Justifications related to recognizing structures	Justifications related to enacting solution steps	Justifications related to planning and evaluation	Total justifications
CG	3.077 (3.378)	4.539 (3.017)[a, b]	3.077 (3.068)[a, b]	11.077 (7.643)[a, b]
EG_{comp}	1.846 (1.281)	9.308 (5.040)[b] *	7.385 (3.176)[b] **	18.923 (6.383)[b] *
$EG_{comp+AT}$	3.462 (2.727)	13.308 (9.446)[a] *	6.769 (2.862)[a] *	24.769 (12.969)[a] *
Effect size d CG vs. EG_{comp}	−0.482	1.148	1.380	1.114
Effect size d CG vs. $EG_{comp+AT}$	0.125	1.251	1.244	1.286
Effect size d EG_{comp} vs. $EG_{comp+AT}$	0.759	0.528	−0.204	0.572

Note. *$p < .05$, **$p < .01$, ***$p < .001$; [a] and [b] indicate significant differences between two groups; effect sizes were indicated by Cohen's *d*.

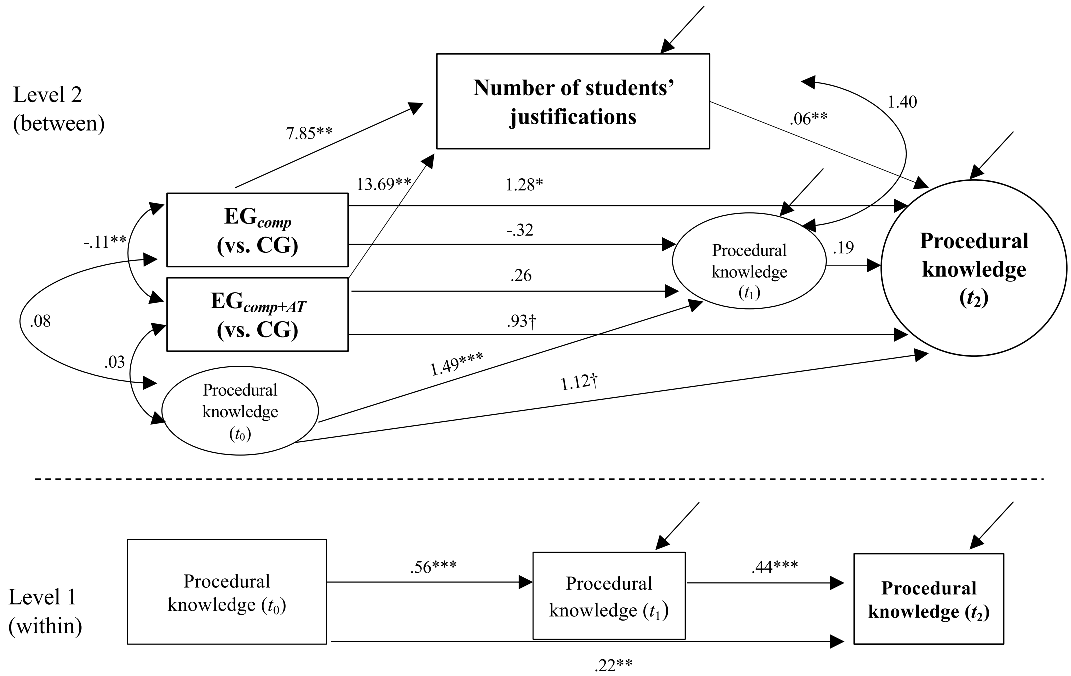

Fig. 3. Multilevel path model for students' justifications as mediator of the effects of both PD programs on procedural knowledge.
Note. †p < .10, *p < . 05, **p < .01, ***p < .001.

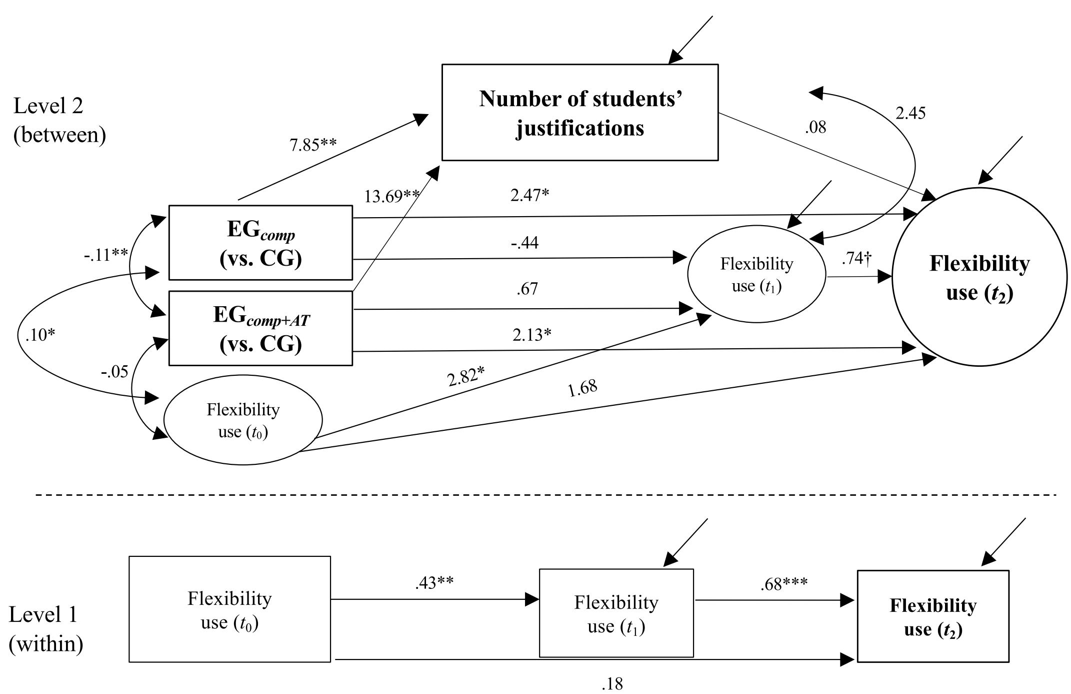

Fig. 4. Multilevel path model for students' justifications as mediator of the effects of both PD programs on flexibility use.
Note. †p < .10, *p < . 05, **p < .01, ***p < .001.

= 4.011, p = .001; Figs. 3–6). The positive effect of justifications on procedural and conceptual knowledge was significant. The effect of justifications on flexibility knowledge was not significant. The effect of justifications on flexibility use was also not significant. We found significant indirect effects for procedural knowledge and conceptual knowledge only in the EG$_{comp+AT}$ (for an overview of all indirect effects, see Table 4). The indirect effect (a_2-path x b-path) on procedural knowledge was significant for the EG$_{comp+AT}$ ($B = 0.848$, $SE = 0.406$, $p =$ .037), but the indirect effect (a_1-path x b-path) on procedural knowledge was not significant for the EG$_{comp}$ ($B = 0.486$, $SE = 0.253$, $p = .055$). A similar pattern emerged for conceptual knowledge: the indirect effect (a_2-path x b-path) on conceptual knowledge was significant for the EG$_{comp+AT}$ ($B = 0.769$, $SE = 0.363$, $p = .034$), but this was not the case for the EG$_{comp}$ ($B = 0.441$, $SE = 0.235$, $p = .061$). The other indirect effects on flexibility use and flexibility knowledge were not significant for either of the EGs in comparison to the CG (flexibility use: EG$_{comp+AT}$ ($B = 1.036$, $SE = 0.796$, $p = .193$); EG$_{comp}$ ($B = 0.594$, $SE = 0.482$, $p =$ .218); flexibility knowledge: EG$_{comp+AT}$ ($B = 1.002$, $SE = 0.702$, $p =$

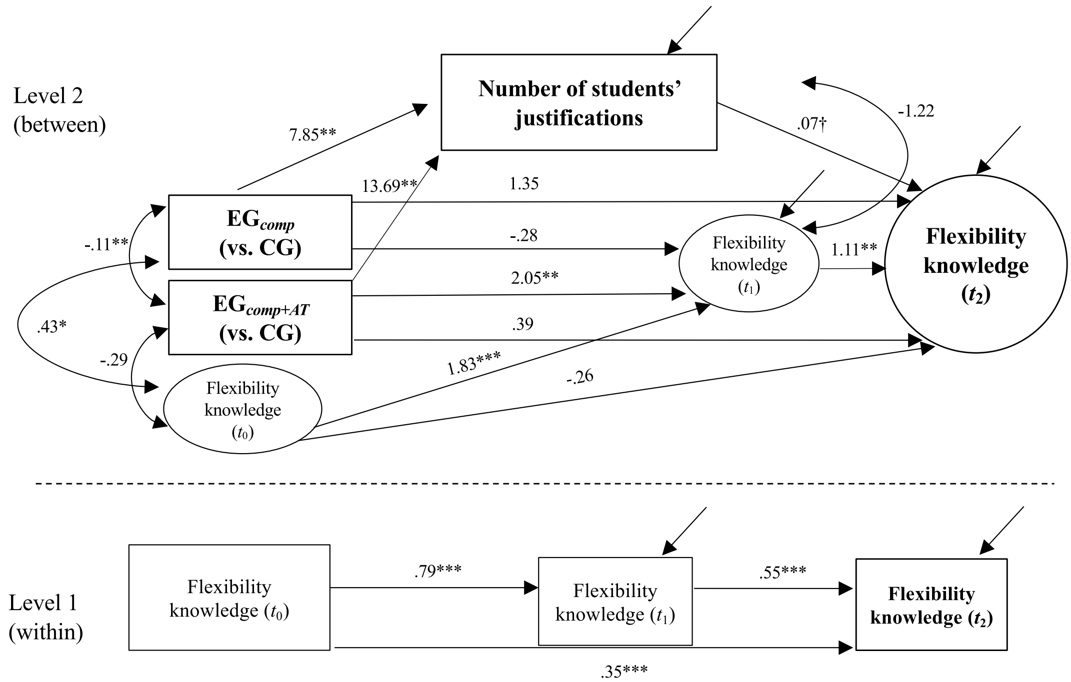

Fig. 5. Multilevel path model for students' justifications as mediator of the effects of both PD programs on flexibility knowledge.
Note. †p < .10, *p < . 05, **p < .01, ***p < .001.

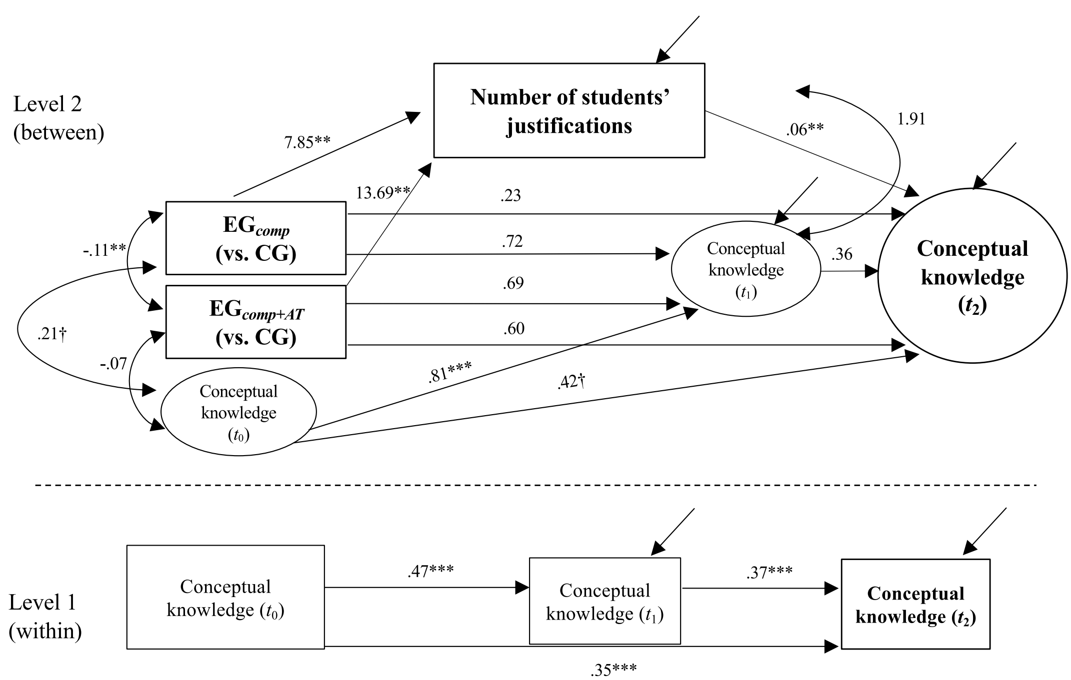

Fig. 6. Multilevel path model for students' justifications as mediator of the effects of both PD programs on conceptual knowledge.
Note. †p< .10, *p < . 05, **p < .01, ***p < .001.

.153); EG$_{comp}$ ($B = 0.575$, $SE = 0.411$, $p = .162$; Table 4). All total effects were significant in our models. The direct effect on procedural knowledge was significant for the EG$_{comp}$ and not significant for the EG$_{comp+AT}$. The direct effects on flexibility use remained significant for both groups. All other direct effects on flexibility knowledge and conceptual knowledge were not significant. Finally, we tested whether the mediation effects of the EGs differ from each other. The results showed that there were no significant differences between the indirect effects of the two EGs on procedural knowledge ($B = -0.362$, $SE = 0.271$, $p = .181$) and conceptual knowledge ($B = -0.328$, $SE = 0.239$, $p = .169$).

For EG$_{comp}$ compared the CG, all indirect effects were not significant and were negligibly small (all *SMD*s $\leq$ .115, all *p*s $\geq$ .055; Cheung, 2009, Table 4). For EG$_{comp+AT}$ compared to the CG, three indirect effects showed small effects (procedural knowledge: $SMD = 0.185$, $p = .037$; flexibility use: $SMD = 0.191$, $p = .193$; conceptual knowledge: $SMD = 0.200$, $p = .034$); only the indirect effect of flexibility knowledge was

Table 4
Indirect effects of both EGs (in comparison to the CG) via justifications on all achievement sub-scales.

	B	*SMD*[a]	*SE*	*p*
EG_{comp}				
Procedural knowledge	.486	.106	.253	.055
Flexibility use	.594	.110	.482	.218
Flexibility knowledge	.575	.075	.411	.162
Conceptual knowledge	.441	.115	.235	.061
$EG_{comp+AT}$				
Procedural knowledge	.848	.185	.406	.037
Flexibility use	1.036	.191	.796	.193
Flexibility knowledge	1.002	.130	.702	.153
Conceptual knowledge	.769	.200	.363	.034

Notes. As recommended for studies with two treatment groups and multiple outcome variables in complex models (Cheung, 2018), the unstandardized scores of each outcome were standardized by the *SD* of each outcome. [a] The standardized mean difference (*SMD*) in our study can be interpreted using the effect size benchmarks from Cohen (1988), adapted for indirect effects (i.e., small effect: $SMD = 0.14$, medium effect: $SMD = 0.36$, and large effect: $SMD = 0.51$; Cheung, 2009, p. 432).

negligible ($SMD = 0.130$, $p = .153$). In sum, for the $EG_{comp+AT}$ (compared to the CG), the indirect effects on procedural knowledge and conceptual knowledge were significant and indicated small effects.

4. Discussion

Our study, based on 39 teachers and 739 9th and 10th grade students, is one of few studies investigating PD effects not only on students' justifications during classroom discourse but also on achievement as a desirable long-term outcome (Desimone, 2009). Both PD programs led to more students justifications in classroom discourse during the teaching unit on quadratic equation solving in comparison to the CG. The number of students justifications between both PD programs did not differ significantly. With multilevel path models, we revealed that the number of students' justifications significantly mediated the effect of a PD focusing on comparing and AT on students' procedural and conceptual knowledge, but not on strategy flexibility. For the PD focusing on comparing only, no mediation effects were significant. The mediation effects of both PD programs did not differ significantly.

Our mediation effects add to the few studies exploring the effects of comparing solution methods in real classrooms (Durkin et al., 2021; Rüede et al., in press; Star et al., 2015). We found evidence for justifications as a mediator in the $EG_{comp+AT}$: the more students in this group generated subject matter justifications during classroom discussions, the higher their procedural and conceptual knowledge was in the posttest. The observed mediation effects in the $EG_{comp+AT}$ on procedural and conceptual knowledge also contribute to productive classroom discourse research in mathematics classrooms. First, many studies on teacher practices in mathematics lessons explore a small number of classes and analyze dialogical patterns qualitatively (e.g., Franke et al., 2009; Kazemi & Stipek, 2009). In contrast, our study used an experimental design to investigate the underlying classroom discourse processes beyond the direct PD effects on student achievement. Second, our study is one of few studies analyzing students' subject matter justifications instead of explanations (cf. Kazemi & Stipek, 2009). It should be noted, however, that the use of the terms explanation and justification strongly depends on their definitions and therefore there may be some overlap in these definitions in different studies (cf. Mata-Pereira & da Ponte, 2017).

4.1. PD effects on students justifications

The PD effects on students justifications supported our *Hypothesis 1a* and *1b* that students in both PD programs would significantly generate more students justifications than the CG. We found large effects of both PD programs on students justifications (EG_{comp}: $d = 1.114$; $EG_{comp+AT}$: $d = 1.286$, Table 3). The effect of our PD programs on students justifications is consistent with previous research on comparing and justifications that postulate a positive influence of the use of comparing tasks with prompts on students justifications (e.g., Rittle-Johnson & Star, 2007; Stein et al., 1996; see 1.1.2). The two PD programs, however, did not differ significantly with regard to the number of students justifications. This is contrary to our *Hypothesis 1c*. It might be that teachers in the $EG_{comp+AT}$ let students build more on each other's ideas in comparison to teachers in the EG_{comp} who might have worked more closely with the prompts in the worked-out examples of the comparing tasks only—leading to more discussion but similar numbers of students justifications. Supporting this, the analysis of student turns shows that there was more student engagement during classroom discourse in the $EG_{comp+AT}$ than in the EG_{comp} (Table 2).

4.2. Mediation effects on procedural knowledge, conceptual knowledge, and strategy flexibility

In the following section, we discuss the mediation effects on procedural and conceptual knowledge for each EG. The non-significant mediation effects on strategy flexibility for both EGs will then be discussed together.

4.2.1. Mediation effects on procedural and conceptual knowledge

All mediation effects for the EG_{comp} compared to the CG were not significant and were negligibly small (*SMD*s $< .14$), contradicting *Hypothesis 2a* that predicted significant mediation effects for all achievement outcomes. We discuss possible explanations for the non-significant mediation effects of the EG_{comp} on procedural and conceptual knowledge. We found that EG_{comp} ($M = 211$, $SD = 84$) descriptively showed more student turns during classroom discourse than the CG ($M = 118$, $SD = 73$; Table 2). Thus, students in the EG_{comp} showed a trend to be more engaged in the classroom discussion than students in the CG. The non-significant trend in favor of student engagement in the EG_{comp} may explain why the mediation effects on procedural and conceptual knowledge for this group missed the significance level of .05. The findings for the EG_{comp} are partially in line with previous comparing research. A recent study by Durkin et al. (2021) found no PD effects on procedural or conceptual knowledge from comparing after letting students discuss in pairs and in the classroom. There were, however, significant effects on students' strategy flexibility and overall achievement score. These results should be interpreted with caution, however, as the knowledge subscales had a low internal consistency (α ranging from .45 to .60), likely due to the low number of items per subscale (Durkin et al., 2021). It also should be noted that Durkin et al. (2021) investigated PD effects on student outcomes, not the mediating effect of justifications. Nevertheless, they developed study materials to support classroom and pair discussions while using worked-out examples to compare solution methods, which likely influenced students justifications.

Our significant mediation effects for the $EG_{comp+AT}$ on procedural and conceptual knowledge are partially in line with *Hypothesis 2b*, predicting significant mediation effects in the $EG_{comp+AT}$ for all achievement outcomes. The mediation effects for the $EG_{comp+AT}$ were small (procedural knowledge $SMD = 0.185$; conceptual knowledge $SMD = 0.200$; Cheung, 2009). Based on our intervention fidelity data, we found that both EGs used more comparing tasks with worked-out examples than the CG (Table 2). To take advantage of such comparing tasks more broadly as opportunities for learning in mathematics (Verschaffel et al., 2009), a deliberate focus on classroom discourse (e.g., through AT) seems necessary. AT specifically aims to generate and work with students justifications, which appears to additionally foster procedural and conceptual knowledge in the $EG_{comp+AT}$. Moreover, the number of student turns was significantly higher in the $EG_{comp+AT}$ ($M = 285$, $SD = 123$) than in the CG ($M = 118$, $SD = 73$, Table 2). The number of student turns can at least partly explain the significant mediation effects of students justifications on procedural and conceptual knowledge for $EG_{comp+AT}$. Drawing on these results and in line with the theoretical assumptions of AT (Michaels et al., 2008), student engagement during comparing solution methods can not only lead to student-generated justifications, but also to students building more on other students' ideas during discussions, which can facilitate linking of procedures and concepts. The lack of a deliberate focus on initiating and orchestrating student engagement in discourse related to comparing solutions in the EG_{comp} may have resulted in fewer justifications (cf. student turns; Table 2).

Our mediation effects on procedural and conceptual knowledge in the $EG_{comp+AT}$ are consistent with previous findings. Our result on conceptual knowledge is in accordance with Kazemi and Stipek (2009): They found that conceptual thinking is promoted when teachers press their students to bring up "explanations with mathematical arguments" (which we defined as justifications) rather than procedural descriptions, make connections between mathematical strategies, use errors to reconceptualize problems, and provide opportunities for individual accountability to come to a consensus about a solution problem. Our findings for the $EG_{comp+AT}$ are also in accordance with studies showing that contributing justifications connecting procedures and concepts in classroom discussions increase students' procedural and conceptual knowledge (Legesse, Luneta, & Ejigu, 2020; Mata-Pereira & da Ponte, 2017).

4.2.2. Mediation effects on strategy flexibility

Students justifications were not a significant mediator of strategy flexibility for the EG_{comp} and $EG_{comp+AT}$. These results were unexpected (*Hypothesis 2a* and *2b*). There are potential explanations as to why we did not find mediation effects on strategy flexibility in both EGs. We did find a salient difference: Both EGs used comparing tasks with worked-out examples significantly more than the CG (EG_{comp}: $M = 4.85$, $SD = 3.31$; $EG_{comp+AT}$: $M = 3.62$, $SD = 4.11$; CG: $M = 0.08$, $SD = 0.28$; Table 2). Thus, it seems that teachers using comparing tasks with worked-out examples create learning opportunities that are almost never provided in the CG classrooms. Presenting students with worked-out examples of comparing tasks seems to foster flexibility in solving such equations – without students justifications playing a major role. Using worked-out examples in comparing tasks has been shown to directly support student gains in strategy flexibility (Durkin et al., 2017; Rittle-Johnson et al., 2012). Little is still known, however, about comparing tasks as a mediator of PD effects and student outcomes. It would be interesting for future studies to investigate comparing tasks (and their prompts) as mediators of the PD effects on strategy flexibility.

The non-significant mediation effects of both EGs on strategy flexibility might be explained by further considering the content of the student turns. In our classroom discourse coding of justifications focusing on why solutions steps, procedures, and strategies work, we did not investigate the focus of students' explanations while comparing. Rittle-Johnson and Star (2009) coded the "explanation characteristics" of students' explanations, such as comparing efficiency or whether or not students referenced multiple methods. The authors found that in the comparing tasks condition (compared to other conditions), students in peer discussions focused more on the solution methods, referenced multiple methods, and were more likely to evaluate the efficiency of methods. Students in this condition also outperformed students from other conditions on strategy flexibility. However, the relative importance of these student explanation characteristics for strategy flexibility has not yet been explored. It seems likely that student explanation characteristics may also be a potential mediator of PD effects on strategy flexibility. Future research should explore the different explanation characteristics suggested by Rittle-Johnson and Star (2009) to gain more insight into the mediating variables affecting strategy flexibility.

4.3. Differences between the mediation effects in both EGs

Contrary to our expectations in *Hypothesis 2c*, we did not find significant differences in the size of the mediation effects between the two EGs on procedural and conceptual knowledge. This could be due to low statistical power (13 teachers per group) for the mediation effects given our two-level design with two EGs. Hence, we recommend future studies strive for a larger sample size to increase the statistical power. Even though the indirect effects of both EGs did not differ significantly, we would like to draw attention to the effect size of the indirect effects. For the $EG_{comp+AT}$, the significant indirect effects on procedural and conceptual knowledge indicated small effects ($SMD = 185$; $SMD = 0.200$), whereas the non-significant indirect effects for the same outcomes for the EG_{comp} indicated negligible effects (SMDs $< .14$; Cheung, 2009, Table 4). However, even the small effect size we observed based on only four videotaped lessons may have large practical relevance, as small indirect effects can accumulate to create larger effects over a school year (Funder & Ozer, 2019). The size of indirect effects in the $EG_{comp+AT}$ (for procedural knowledge: $SMD = 0.185$; for conceptual knowledge: $SMD = 0.200$) are nearly twice as large as the indirect effects of the EG_{comp} (for procedural knowledge: $SMD = 0.106$; for conceptual knowledge: $SMD = 0.115$). Furthermore, the $EG_{comp+AT}$ showed clear significant indirect effects compared to the CG. Thus, the $EG_{comp+AT}$ was successful (compared to the CG) in increasing students' subject matter justifications, which improved the procedural and conceptual knowledge of students.

4.4. Limitations

The reliability values for the conceptual knowledge scales across different time points (and for procedural knowledge at baseline) were rather low. The low reliability for the conceptual knowledge scale might have resulted from the slightly different types of items used for the knowledge scales as compared to the questions students usually had to solve in school. Moreover, the low reliability of procedural knowledge at baseline may be explained by a floor effect because all students had limited knowledge regarding the topic of solving quadratic equations at the beginning of the school year. We investigated the effects of two PD programs that focused either on comparing alone or comparing along with AT, while the effect of AT alone was not investigated. Future studies should explore a separate EG including also a group focusing on the orchestration of productive discourse for solving equations based on the approach of AT only. Finally, our mediation results are related to classroom discourse with respect to solving quadratic equations and, therefore, limited to this topic. It would be interesting to explore students' justifications in other mathematical topics such as solving linear equations, as none of the several studies demonstrating effects of comparing solution methods on achievement explored mediation effects (Durkin et al., 2017).

5. Conclusion

The results of our multilevel path models based on 39 teachers and 739 students revealed that the subject matter justifications during classroom discourse significantly mediated the effect of the combined PD comparing+AT in comparison to the CG on procedural and conceptual knowledge, but not on strategy flexibility. These findings point to the importance for teachers learning to orchestrate productive classroom discourse regarding subject matter content while using comparing solution methods to foster not only students' subject matter justifications in classroom discussions, but also their procedural and conceptual knowledge in mathematics. No significant mediation effects for students' subject matter justifications were found for the PD with comparing only.

Authorship contribution statement

Sog Yee Mok: Conceptualization, Methodology, Formal analysis, Writing - Original Draft, Writing - Review & Editing, Visualization, Project administration. **Christian S. Hämmerle**: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Data Curation, Resources, Writing - Review & Editing. **Christian Rüede**: Conceptualization, Methodology, Validation, Investigation, Resources, Writing - Review & Editing, Funding acquisition. **Fritz C. Staub**: Conceptualization, Methodology, Validation, Investigation, Resources, Writing - Review & Editing, Funding acquisition.

Funding

Swiss National Science Foundation Grant No. 100019_162686/1.

Declaration of competing interest

None.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.learninstruc.2022.101668.

References

Asterhan, C. S., & Schwarz, B. B. (2016). Argumentation for learning: Well-trodden paths and unexplored territories. *Educational Psychologist, 51*(2), 164–187.
Beck, I. L., & McKeown, M. B. (2006). *Improving comprehension with questioning the author: A fresh and expanded view of a powerful approach.* New York: Scholastic Inc.
Booth, J. L., Barbieri, C., Eyer, F., & Paré-Blagoev, E. J. (2014). Persistent and pernicious errors in algebraic problem solving. *Journal of Problem Solving, 7*, 11–23.
Chen, G., Chan, C. K. K., Chan, K. K. H., Clarke, S. N., & Resnick, L. B. (2020). Efficacy of video-based teacher professional development for increasing classroom discourse and student learning. *The Journal of the Learning Sciences*, 1–39.
Cheung, M. W. L. (2009). Comparison of methods for constructing confidence intervals of standardized indirect effects. *Behavior Research Methods, 41*(2), 425–438. https://doi.org/10.3758/BRM.41.2.425
Cheung, M. W. L. (2018). Computing multivariate effect sizes and their sampling covariance matrices with structural equation modeling: Theory, examples, and computer simulations. *Frontiers in Psychology, 9*. https://doi.org/10.3389/fpsyg.2018.01387
Cooper, G., & Sweller, J. (1987). Effects of schema acquisition and rule automation on mathematical problem-solving transfer. *Journal of Educational Psychology, 79*, 347–362.
Desimone, L. M. (2009). Improving impact studies of teachers' professional development: Toward better conceptualizations and measures. *Educational Researcher, 38*, 181–199.
Durkin, K., Rittle-Johnson, B., Star, J. R., & Loehr, A. (2021). Comparing and discussing multiple strategies: An approach to improving algebra instruction. *The Journal of Experimental Education*, 1–19.
Durkin, K., Star, J. R., & Rittle-Johnson, B. (2017). Using comparison of multiple strategies in the mathematics classroom: Lessons learned and next steps. *ZDM Mathematics Education, 49*, 585–597.
Enders, C. K. (2010). *Applied missing data analysis*. New York: Guilford press.
Franke, M. L., Webb, N. M., Chan, A. G., Ing, M., Freund, D., & Battey, D. (2009). Teacher questioning to elicit students' mathematical thinking in elementary school classrooms. *Journal of Teacher Education, 60*(4), 380–392.
Fritz, M. S., & MacKinnon, D. P. (2007). Required sample size to detect the mediated effect. *Psychological Science, 18*(3), 233–239.
Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science, 2*(2), 156–168. https://doi.org/10.1177/2515245919847202
Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1–55.
Inglis, M., & Mejia-Ramos, J. P. (2008). How persuaded are you? A typology of responses. *Research in Mathematics Education, 10*(2), 119–133.
Ing, M., Webb, N. M., Franke, M. L., Turrou, A. C., Wong, J., Shin, N., et al. (2015). Student participation in elementary mathematics classrooms: The missing link between teacher practices and student achievement? *Educational Studies in Mathematics, 90*(3), 341–356.
Kazemi, E., & Stipek, D. (2009). Promoting conceptual thinking in four upper-elementary mathematics classrooms. *Journal of Education, 189*(1–2), 123–137. https://doi.org/10.1177/0022057409189001-209
Kennedy, M. M. (2016). How does professional development improve teaching? *Review of Educational Research, 86*(4), 945–980. https://doi.org/10.3102/0034654315626800
Kieran, C. (1989). The early learning of algebra: A structural perspective. In S. Wagner, & C. Kieran (Eds.), *Research Issues in the Learning and Teaching of algebra (S. 33–56)*. Hillsdale: Erlbaum.
Kim, M.-Y., & Wilkinson, I. A. G. (2019). What is dialogic teaching? Constructing, deconstructing, and reconstructing a pedagogy of classroom talk. *Learning, Culture and Social Interaction, 21*, 70–86.
Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174. https://doi.org/10.2307/2529310
Legesse, M., Luneta, K., & Ejigu, T. (2020). Analyzing the effects of mathematical discourse-based instruction on eleventh-grade students' procedural and conceptual understanding of probability and statistics. *Studies In Educational Evaluation, 67*, Article 100918.
Linchevski, L., & Livneh, D. (1999). Structure sense: The relationship between algebraic and numerical contexts. *Educational Studies in Mathematics, 40*, 173–199.
Litke, E. (2020). Instructional practice in algebra: Building from existing practices to inform an incremental improvement approach. *Teaching and Teacher Education, 91*, Article 103030.
Lüdtke, O., Marsh, H. W., Robitzsch, A., & Trautwein, U. (2011). A 2× 2 taxonomy of multilevel latent contextual models: Accuracy–bias trade-offs in full and partial error correction models. *Psychological Methods, 16*(4), 444–467.
Mata-Pereira, J., & da Ponte, J.-P. (2017). Enhancing students' mathematical reasoning in the classroom: Teacher actions facilitating generalization and justification. *Educational Studies in Mathematics, 96*(2), 169–186.
Matsumura, L. C., Garnier, H. E., & Spybrook, J. (2013). Literacy coaching to improve student reading achievement: A multi-level mediation model. *Learning and Instruction, 25*, 35–48. https://doi.org/10.1016/j.learninstruc.2012.11.001
Michaels, S., O'Connor, C., & Resnick, L. B. (2008). Deliberative discourse idealized and realized: Accountable Talk in the classroom and in civic life. *Studies in Philosophy and Education, 27*(4), 283–297.
Michaels, S., O'Connor, C., Hall, M. W., & Resnick, L. B. (2010). *Accountable talk sourcebook*. Pittsburgh: University of Pittsburgh.
National Council of Teachers of Mathematics. (2006). *Curriculum focal points for prekindergarten through grade 8 mathematics* (Reston, Vol. A).
O'Connor, C., Michaels, S., Chapin, S., & Harbaugh, A. G. (2017). The silent and the vocal: Participation and learning in whole-class discussion. *Learning and Instruction, 48*, 5–13.
Pehmer, A.-K., Gröschner, A., & Seidel, T. (2015). Fostering and scaffolding student engagement in productive classroom discourse: Teachers' practice changes and reflections in light of teacher professional development. *Learning, Culture and Social Interaction, 7*, 12–27. https://doi.org/10.1016/j.lcsi.2015.05.001
Pimentel, D. S., & McNeill, K. L. (2013). Conducting talk in secondary science classrooms: Investigating instructional moves and teachers' beliefs. *Science Education, 97*(3), 367–394. https://doi.org/10.1002/sce.21061
Preacher, K. J., & Kelley, K. (2011). Effect size measures for mediation models: Quantitative strategies for communicating indirect effects. *Psychological Methods, 16*(2), 93–115.
Resnick, L. B., Asterhan, C. S. C., Clarke, S., & Schantz, F. (2018). Next generation research in dialogic learning. In G. E. Hall, L. F. Quinn, & D. M. Gollnick (Eds.), *Wiley handbook of teaching and learning.* Wiley-Blackwell, 338-323.
Resnick, L., Michaels, S., & O'Connor, C. (2010). How (well-structured) talk builds the mind. In D. D. Preiss, & R. J. Sternberg (Eds.), *Innovations in educational psychology* (pp. 163–194). New York: Springer.
Reznitskaya, A., Glina, M., Carolan, B., Michaud, O., Rogers, J., & Sequeira, L. (2012). Examining transfer effects from dialogic discussions to new tasks and contexts. *Contemporary Educational Psychology, 37*(4), 288–306. https://doi.org/10.1016/j.cedpsych.2012.02.003
Rittle-Johnson, B., & Star, J. R. (2007). Does comparing solution methods facilitate conceptual and procedural knowledge? An experimental study on learning to solve equations. *Journal of Educational Psychology, 99*, 561–574.
Rittle-Johnson, B., & Star, J. R. (2009). Compared with what? The effects of different comparisons on conceptual knowledge and procedural flexibility for equation solving. *Journal of Educational Psychology, 101*, 529–544.
Rittle-Johnson, B., Star, J. R., & Durkin, K. (2020). How can cognitive-science research help improve education? The case of comparing multiple strategies to improve mathematics learning and teaching. *Current Directions in Psychological Science, 29*(6), 599–609.

Rittle-Johnson, B., Star, J. R., & Durkin, K. (2012). Developing procedural flexibility: Are novices prepared to learn from comparing procedures? *British Journal of Educational Psychology, 82*, 436–455.

Rüede, C., Mok, S. Y., & Staub, F. C. (in press). Fostering flexibility using comparing solution methods and classroom discussions: Effects of two professional development programs. *Journal for Research in Mathematics Education*.

Rüede, C., & Staub, F. C. (2019). Adaptivität als Kern basaler mathematischer Kompetenzen für allgemeine Studierfähigkeit: Was heisst das in der Algebra? [Adaptivity as the core of basic mathematical competencies for college readiness: What does it mean in algebra?]. In D. Holtsch, & M. Oepke (Eds.), *Lehren und Lernen auf der Sekundarstufe II* (pp. 188–198). Bern: hep.

Schneider, M., Rittle-Johnson, B., & Star, J. R. (2011). Relations among conceptual knowledge, procedural knowledge, and procedural flexibility in two samples differing in prior knowledge. *Developmental Psychology, 47*, 1525–1538.

Schoenfeld, A. H. (1987). What's all the fuss about metacognition? In A. H. Schoenfeld (Ed.), *Cognitive science and mathematics education* (pp. 189–216). Hillsdale: Erlbaum.

Sedova, K., Sedlacek, M., & Svaricek, R. (2016). Teacher professional development as a means of transforming student classroom talk. *Teaching and Teacher Education, 57*, 14–25.

Sedova, K., Sedlacek, M., Svaricek, R., Majcik, M., Navratilova, J., Drexlerova, A., … Salamounova, Z. (2019). Do those who talk more learn more? The relationship between student classroom talk and student achievement. *Learning and Instruction, 63*, Article 101217.

Sowder, L., & Harel, G. (1998). Types of students' justifications. *Mathematics Teacher, 91* (8), 670–675.

Star, J. R., Pollack, C., Durkin, K., Rittle-Johnson, B., Lynch, K., Newton, K., et al. (2015). Learning from comparison in algebra. *Contemporary Educational Psychology, 40*, 41–54.

Stein, M. K., Grover, B. W., & Henningsen, M. (1996). Building student capacity for mathematical thinking and reasoning: An analysis of mathematical tasks used in reform classrooms. *American Educational Research Journal, 33*, 455–488.

van der Veen, C., de Mey, L., van Kruistum, C., & van Oers, B. (2017). The effect of productive classroom talk and metacommunication on young children's oral communicative competence and subject matter knowledge: An intervention study in early childhood education. *Learning and Instruction, 48*, 14–22. https://doi.org/10.1016/j.learninstruc.2016.06.001

Verschaffel, L., Luwel, K., Torbeyns, J., & Van Dooren, W. (2009). Conceptualizing, investigating, and enhancing adaptive expertise in elementary mathematics education. *European Journal of Psychology of Education, 24*(3), 335–359.

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

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

Webb, N. M., Franke, M. L., Ing, M., Chan, A., De, T., Freund, D., et al. (2008). The role of teacher instructional practices in student collaboration. *Contemporary Educational Psychology, 33*(3), 360–381.

Webb, N. M., Franke, M. L., Ing, M., Wong, J., Fernandez, C. H., Shin, N., et al. (2014). Engaging with others' mathematical ideas: Interrelationships among student participation, teachers' instructional practices, and learning. *International Journal of Educational Research, 63*, 79–93.