

Fostering conceptual knowledge of procedures by explicating structures in multiple representations: Classwise randomized controlled trial for the case of unit conversion

Sofia Peters ^{1*} , Susanne Prediger ^{1,2} , Juliane Weiss ² 

¹ TU Dortmund University, Vogelpothsweg 87, 44227 Dortmund, GERMANY

² IPN Leibniz Institute for Science and Mathematics Education, Berlin, GERMANY

*Corresponding Author: sofia.peters@math.tu-dortmund.de

Citation: Peters, S., Prediger, S., & Weiss, J. (2026). Fostering conceptual knowledge of procedures by explicating structures in multiple representations: Classwise randomized controlled trial for the case of unit conversion. *International Electronic Journal of Mathematics Education*, 21(4), em0897. <https://doi.org/10.29333/iejme/19642>

ARTICLE INFO

Received: 14 Jul. 2026

Accepted: 24 Sep. 2026

ABSTRACT

Instructional approaches with multiple representations and rich discourse practices can effectively foster conceptual knowledge of mathematical procedures. Yet, qualitative studies indicate that supporting students in explicitly articulating the underlying mathematical structures might even increase learning gains. This hypothesis is tested in a classwise randomized controlled trial (with $N = 434$ students) for the example topic unit conversion in digitally assisted teaching-learning environments. Both intervention groups (using multiple representations and rich discourse practices) had significantly higher learning gains than the control group in conceptual knowledge and content-related language proficiency. The structure-explicating intervention showed significant additional conceptual and language-related learning gains. These findings contribute to theorizing on principles of conceptually focused instruction and the high relevance of engaging students in explicating structures.

Keywords: unit conversion, procedural and conceptual knowledge, controlled trial, digitally-assisted teaching, learning environment, structure explication

INTRODUCTION: MULTIPLE REPRESENTATIONS, RICH DISCOURSE PRACTICES, AND EXPLICATING STRUCTURES FOR FOSTERING CONCEPTUAL KNOWLEDGE OF PROCEDURES

It is widely accepted that when students learn new mathematical procedures (such as multi-digit multiplication, expanding fractions, or unit conversion), the learning goals should not only comprise procedural but also conceptual knowledge (Freudenthal, 1981; Kilpatrick et al., 2001). While *procedural knowledge* refers to conducting a procedure fluently, *conceptual knowledge for procedures* refers to what the procedure means and why it can be conducted in the learned way (Hiebert & Lefevre, 1986). It usually builds upon conceptual knowledge for the concepts and structures involved (Freudenthal, 1981; Fuson, 2019). For many decades, such conceptual knowledge for concepts and procedures has been fostered by the *design principle of connecting multiple representations* for constructing meanings (Cramer et al., 2002; NCTM, 2014; Prediger, 2022).

In the last two decades, the research on language-responsive mathematics instruction (Erath et al., 2021) added a second design principle for effectively fostering conceptual knowledge: *engaging students in rich discourse practices* (Moschkovich, 2015) while providing scaffolding for the content-related language (Gibbons, 2002; Prediger et al., 2022; Zahner et al., 2021). Based upon many years of insightful qualitative research on this design principle (Walshaw & Anthony, 2008), also quantitative evidence for its efficacy was given in controlled trials for arithmetical procedures (surveyed in Erath et al., 2021).

Qualitative case studies on learning processes with multiple representations and rich discourse practices revealed another principle that might further augment the learning gains by deepening students' learning experience: Many students tend to conduct surface translations between representations and apply quasi-empirical modes of dealing with visual representations without focusing the underlying mathematical structures (Fuson et al., 1997; Fuson & Briars, 1990; Steinbring, 2005). Therefore, it might be essential to *engage students in explicitly articulating the underlying structures* in order to deepen the conceptual knowledge (Fuson & Briars, 1990; Peters & Prediger, 2026; Tondorf & Prediger, 2022). So far, however, this more specific principle of engaging students in explicitly articulating structures was treated as a subprinciple of the principle engaging students in rich discourse practices and not investigated separately with respect to its additional effects in controlled trials.

In this paper, we present a controlled efficacy trial that transfers the well-established design principles of connecting multiple representations and engaging students in rich discourse practices from arithmetical procedures to a procedure that is still under-researched, the *conversion of mass units* (Cheeseman et al., 2017). While interventions for conceptual knowledge of length, area and volume units have been designed and shown to be effective (Battista, 2007; Smith & Barrett, 2017), students' conceptual knowledge of converting measurement units was only captured in assessment studies (Tan Sisman & Aksu, 2016) but hardly fostered in interventions with the design principles established for arithmetical procedures.

The controlled trial contributes to three strands of research and theorizing: (a) It transfers design principles from conceptually focused interventions on arithmetic procedures to the under-researched content of mass unit conversion. This contributes to expanding the scope of effective conceptually focused instructional designs. (b) It disentangles potential effects with respect to three different but interrelated learning goals, procedural knowledge, conceptual knowledge and the content-related language. This contributes to nuanced findings on different knowledge components. (c) It investigates additional effects of a particular subprinciple, supporting the explication of underlying structures, which was found relevant in qualitative studies but has not yet been studied separately with respect to measurable additional effects. In sum, the study contributes to theorizing on principles of conceptually focused instruction on mass unit conversion and beyond. For this, it treats the following overall research question which will be refined after the theory section:

RQ: To what extent can fifth graders' mass unit conversion knowledge be fostered in interventions using multiple representations and rich discourse practices, and is there an additional effect of supporting the explication of underlying structures?

In the following, we present the current state of research on mass unit conversion and design principles for fostering understanding of mass unit conversion and introduce the instructional design of the digitally assisted teaching-learning environment in view (Section *Theoretical Background*). The methodological framework of the randomized controlled trial (Section *Methods of the Randomized Controlled Trial*), the empirical results (Section *Results*), and a discussion (Section *Discussion and Conclusion*) are presented.

THEORETICAL BACKGROUND: DESIGN PRINCIPLES FOR FOSTERING CONCEPTUAL KNOWLEDGE OF PROCEDURES AND THEIR SUBSTANTIATION FOR MASS UNIT CONVERSION

Three Intertwined Types of Knowledge for Arithmetical Procedures

In their seminal paper, Hiebert and Lefevre (1986) distinguished conceptual knowledge as knowledge of concepts which are rich in connections between visual, verbal and symbolic representations (e.g., the meaning of fractions as part-whole relationships) from procedural knowledge as knowledge of procedures in symbolic representations (e.g., the procedure of how to expand fractions), and emphasized that both should be intertwined. Since then, a rich theoretically and empirically grounded discourse disentangled both knowledge types in multiple ways, depending on the topics in view (e.g., summarized in Rittle-Johnson & Schneider, 2015) and extended conceptual knowledge from concepts to procedures: In this paper, we call a *mathematical procedure* a step-by-step prescription of how to complete a certain task by symbolic transformations (e.g., expand the fraction by multiplying denominator and numerator with the same factor). We call *procedural knowledge* the skill to execute the procedure fluently and without error, and call *conceptual knowledge* of a procedure the understanding of *what* the procedure *means* (e.g. for expanding fractions, it transforms a fraction into another one that describes the same part-whole relationship) and *why* the procedure works (e.g., visualize expanding the fraction $\frac{2}{6}$ in a fraction bar by three by refining each of the 6 pieces into 3 smaller pieces, then the new whole consists of 6 groups of 3 pieces, and we multiply the denominator 6×3). The 2 coloured pieces of the fraction bar are also tripled, so to preserve the part-whole relationship, the numerator must also be tripled (Wessel, 2020). Conceptual knowledge of a procedure is more than of the concepts involved (what does the fraction concept mean?), as it also connects the symbolic transformations to restructuring actions in the visual representation for its justification (Fuson, 2019; Tondorf & Prediger, 2022). While students can hold conceptual knowledge also implicitly (Rittle-Johnson & Schneider, 2015), its articulation also requires *content-related language proficiency*: on the discursive level, it requires the discourse practices of explaining meanings and justifying procedures in the visualization, and on the lexical level, it requires the lexical means to enact the discourse practices (Prediger, 2022).

Design Principles for Fostering Procedural and Conceptual Knowledge and Content-Related Language Proficiency of Arithmetical Procedures

To develop conceptual knowledge of procedures, the *design principle of connecting multiple (visual, verbal, material, and symbolic) representations* (DP1) has often proven effective as it helps to construct the meanings (Cramer et al., 2002; Fuson et al., 1997; NCTM, 2014). A necessary condition for effective learning from multiple representations is that students do not only superficially translate the representations, but deeply connect them by explaining how to find the essential elements in the other representation (Renkl et al., 2013). For procedures, these deeper connections can be fostered by progressive schematization for which the learning trajectories start from students' prior knowledge in material and visual representations or imaginable contexts and progressively develop towards the symbolic representation in which students can discover the procedures (Freudenthal, 1981; Lorange & Rinvold, 2014).

These processes of meaning construction and progressive schematization require substantial communication in classrooms (Walshaw & Anthony, 2008), which is best supported by the *second design principle of engaging students in rich discourse practices* (DP2) such as explaining meanings, justifying the rules or describing how representations are connected (Erath et al., 2021;

Moschkovich, 2015). For being able to participate in these rich discourse practices, students need to develop *content-related language proficiency* on both, the discursive level (knowing what kind of explanation counts as acceptable in mathematics classrooms) and on the lexical level (having the suitable vocabulary and phrases needed to articulate the contributions in the discourse practices, Prediger, 2022). By scaffolding through discursive models and lexical support (e.g. by offered phrases or sentence starters), the development of this language proficiency can be enhanced (Erath et al., 2021; Gibbons, 2002).

Conceptually focused, language-responsive instructional approaches (designed according to these two design principles) have been shown to effectively foster conceptual understanding of arithmetical procedures, in randomized or quasi-experimental trials, e.g., for ordering fractions (Cramer et al., 2002); expanding fractions (Prediger et al., 2022), multiplication (Götze & Baiker, 2021), or determining percentages (Pöhler et al., 2017). These interventions mainly explicitly fostered conceptual knowledge but showed some effects also for the procedural knowledge in view, which then has to be routinized and flexibilized as well (Kilpatrick et al., 2001). These documented effects allow us not to investigate specific interventions on procedural knowledge. While the general efficacy of these two design principles was thereby shown, the qualitative analysis of teaching-learning processes revealed that a more specific subprinciple of engaging students in rich discourse practices might further nurture the learning gains:

Conceptual knowledge of how to justify procedures seems to be developed only by some students whereas others are limited to reading off results from the visual representation in quasi-empirical modes without understanding their backgrounds (Lorange & Rinvold, 2014; Steinbring, 2005) or discovering non-justifiable rules for the symbolic transformation (Glade & Prediger, 2017). Students who develop conceptual knowledge for justifying procedures seem to be those who actively structure and restructure the visual model according to the underlying structures and explicitly articulate how the structures and the restructurings are visible in each of the representations (Glade & Prediger, 2017; Renkl et al., 2013; Tondorf & Prediger, 2022). Structuring is thereby described as an active “mental act of constructing an organization or form for an object or set of objects ... [including] establishing units, establishing relationships between units” (Battista & Clements, 1996, p. 282). Explicating how the structures and restructurings are visible in the representations is thereby an indicator of accomplished conceptual knowledge, but presumably also an approach for enhancing students’ mental act of structuring and restructuring (Tondorf & Prediger, 2022). For the example of fractions, the design of interventions can thereby include scaffolds of *structure-explicating phrases* to express these structures (e.g., “the part of the whole”, “fraction bar is refined in further pieces”) on the lexical level. The content-related language proficiency needed to develop conceptual knowledge can thereby be focused on these structure-explicating discourse practices (discursive level) and structure-explicating phrases (lexical level) to articulate them (Prediger, 2022).

For arithmetical procedures, these underlying structures are, for example, part-whole structures for ordering fractions (Cramer et al., 2002; Prediger et al., 2022), refinement structures for expanding fractions (Fuson, 2019), place-value structures and bundle structures for multi-digit addition (Fuson et al., 1997), or multiplicative bundle structures for deriving multiplication tasks (Götze & Baiker, 2021). Although the relevance of these underlying mathematical structures for understanding the procedures has been identified in assessment studies and design research studies, this *specific subprinciple of engaging students in explicitly articulating structures* (DP2’) has not yet been investigated with respect to its additional effects compared to the two overarching design principles DP1 and DP2. In this paper, we intend to conduct this comparison for the case of an under-researched topic, mass unit conversion.

Three Intertwined Types of Knowledge for Mass Unit Conversion and the Underlying Mathematical Structures

Compared to the rich state of research on conceptual knowledge of arithmetical procedures, much less is known in the topic of measurement units, in particular mass units (Cheeseman et al., 2017; Smith & Barrett, 2017). Conceptual knowledge of length, area, volume and mass units has been successfully disentangled into the big ideas what measurement means (Battista, 2007; Tan Sisman & Aksu, 2016) and the justification of the procedure of determining area and volume by the multiplicative formulas (relying on the bundle structures and the meaning of multiplication as counting in bundles, cf. Battista, 2007). For the procedure of *measurement unit conversion*, in contrast, empirical research is rare and focused mainly on procedural knowledge needed in science education (e.g., Dincer & Osmanoglu, 2018), and some conceptual backgrounds for preservice teachers (Nguyen et al., 2023). We transfer the three intertwined types of knowledge to the procedure of mass unit conversion as follows:

- *Procedural knowledge* refers to the skill to execute the procedure fluently and without errors (e.g., transform 2.3kg into gram).
- *Conceptual knowledge* refers to the understanding of *what* the procedure *means* (2.3kg and 2300g describe the equally heavy mass) and *why* the transformation can be conducted by multiplication with 1000 (because each 1-kilogram block corresponds to a 1000-g block).
- The relevant *content-related language proficiency* to explain the meaning and justify the procedure needs to refer to all relevant underlying structures (Prediger, 2022).

In qualitative analyses of the preceding design research study, three underlying relevant structures were identified for the conversion of mass units and for justifying this procedure (Peters & Prediger, 2026): multiplicative bundle structures (Clark & Kamii, 1996; Götze & Baiker, 2021; Lamon, 1996), place value structures (Fuson et al., 1997; Hiebert & Wearne, 1996; Ross, 1989) and refinement structures in which entities are split into equally sized pieces or coarsened into larger pieces (Wessel, 2020), these refinement structures have been identified as being most critical in students’ learning (Peters & Prediger, 2026).

In this paper, we focus on the less investigated mass units rather than better investigated measures of length and area, as mass units were rarely treated with respect to conceptual knowledge and interestingly complex underlying structures, namely multiplicative bundle structures, place value structures and refinement structures.

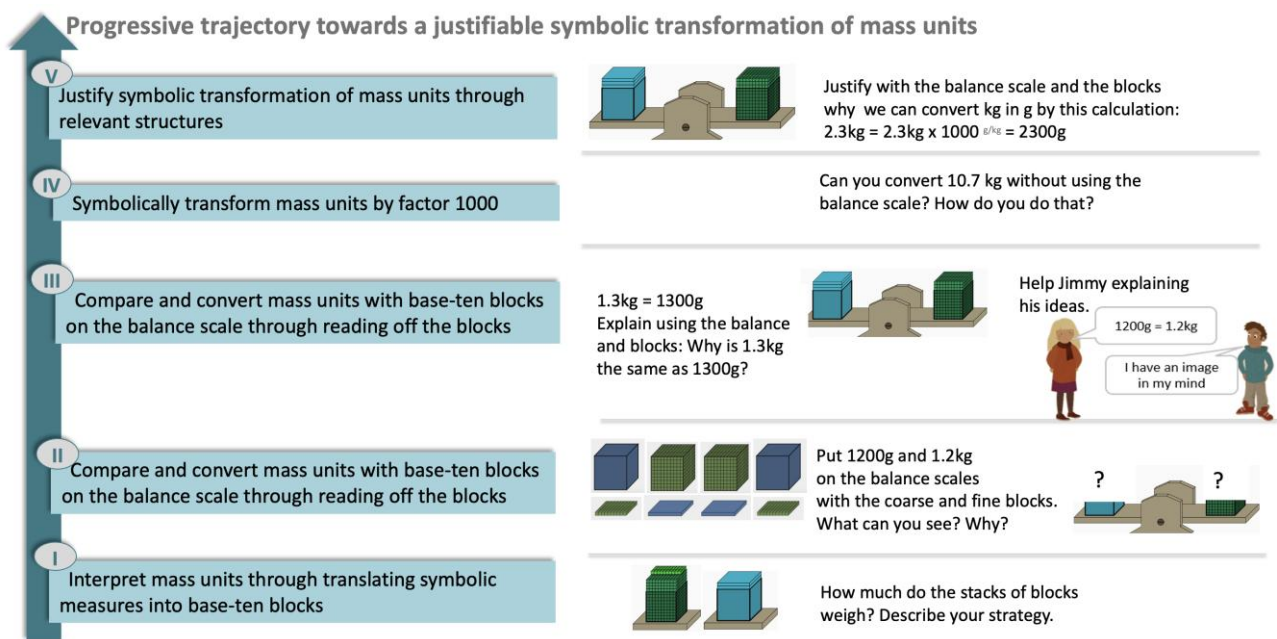


Figure 1. Sequenced learning goals along the progressive learning trajectory with example tasks on each level (Adapted from Peters & Prediger, 2026)

Transferring the Established Design Principles from Arithmetical Procedures to Procedure of Mass Unit Conversion

So far, the two design principles of connecting multiple representations and engaging students in rich discourse practices have not yet been explicitly transferred to the procedure of conversion of measurement units (Cheeseman et al., 2017; Nguyen et al., 2023; Smith & Barrett, 2017). However, Smith and Barrett (2017) briefly report in their survey on teaching measurement about “(a) engaging tasks that involve more than producing or computing a sequence of measures and (b) skillfully led classroom discussion that relates actions to measurement” (p. 378). Based on their empirical findings, they suggest to developing “technological capacity to support learning” (p. 379) and to explore how dynamic visualizations “can be used as productive components of ... measurement teaching” (p. 379).

We followed their call by designing a digitally assisted teaching-learning environment in which students can explore the dynamic visualization of base-ten blocks on a balance scale (Peters & Prediger, 2026). This dynamic visual representation can convey the relational meaning of the equal sign as the *equivalence of two weights* on the balance scale which is first explored in an empirical mode. Connecting the base-ten blocks to the measures supports the *place value structures* (for example that 2300g can be decomposed into its place values, two 1000s in grams and 3 hundreds in grams) and the *multiplicative bundle structures* (1000 in each block and base 10 in each digit for the measures in gram) that are an integral part of place value structures and needed for justifying factors. By offering the coarse (blue) ones-blocks for kilograms and the refined 1000-g-block for grams, the *refinement structures* and their quantitative relation by factor 1000 are addressed in *connecting the multiple representations* (DP1).

In the teaching-learning environment, students’ progressive learning goals are sequenced along the learning trajectory of progressive schematization in **Figure 1**. On each level, students are invited and supported to *engage in rich discourse practices* (DP2) of explaining the connections between the symbolic measures and the visual balance scale representation. They can discover the conversion rule and are invited to justify the procedure of multiplying by 1000 through connecting the representations. Students’ independent work is combined with whole-class discussions for systematizing the gained insights and developed ideas.

To support the students in mastering the high conceptual and discursive demands of two design principles for the case of mass unit conversion, we integrated scaffolds on each level of the learning trajectory. Beyond the dynamic visualization itself (which is a strong visual scaffold for developing the understanding of equally heavy and the variation of stacks of blocks), discursive scaffolds are provided language models how to explain and justify and lexical support for content-related phrases (see **Figure 2**).

In order to study the additional effect of the specific principle of *engaging students in explicitly articulating underlying structures* (DP2’), the digital teaching-learning environment was realized in two versions: the regular intervention **IG** applied the two design principles in the design as well as the teachers’ moderation as presented (second column in **Figure 2**) but without focus on engaging and supporting students in explicitly articulating the underlying structures, whereas the additionally structure-explication intervention **IG+SE** focused the tasks also on connecting multiple representations and engaging in rich discourse practices, and additional focus on engaging the students in explicitly articulating the underlying mathematical structures (third column in **Figure 2**).

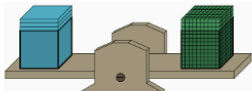
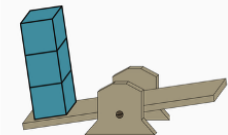
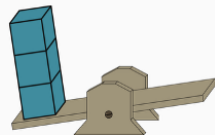
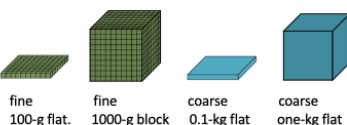
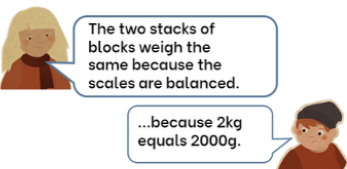
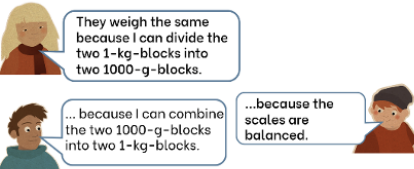
Example task	Regular intervention IG with multiple representations and rich discourse practices	Additionally structure-explicating intervention IG+SE
Design Principle 1: Connecting multiple representations		
	<i>Tasks requesting connecting visual model and symbolic representations</i>	<i>... while using structure-explicating phrases</i>
Example Task 9 from Level III	1.3kg = 1300g. Explain using the balance scale. Why is 1.3kg the same as 1300g? 	1.3kg = 1300g. Explain using the balance scale. Why is 1.3kg the same as 1300g?  Use phrases you have learnt: 1000-gram-block, fine block etc.
Design Principle 2: Engaging students in rich discourse practices (explain and justify connections)		
		Design Principle 2': Engaging students in explicitly articulating underlying structures
High discursive demands	<i>High discursive demands with conceptual and discursive scaffold of visual model supporting the explanation</i>	<i>... and the explicit articulation of the underlying structures by encouraging the use of structure-explicating phrases</i>
Example Task 11 from Level V	Convert 3kg into grams and explain your transformation using the visual model. 	Convert 3kg into grams and explain your transformation using the visual model.  You can explain it like Emma. Use the phrases you have learnt (1000-gram block etc.)
Lexical scaffolds	<i>Lexical scaffolds for explicating structures underlying model shown once on board</i>	<i>... are developed together with the refinement structures, e.g., based on everyday objects</i>
Example Task 2 from Level I	You can use these phrases below if you like. 	Which objects are fine? Which are coarsely split? 
Discursive scaffolds	<i>Modeling formal language use in discourse practice "describing visual model" by fictitious students</i>	<i>Modeling "explicating underlying structures" in discourse practice of fictitious students</i>
Example Task 10 from Level III	Explain in your own words. Which explanations do the two use? 	Explain in your own words. Which explanations do the three students use? 

Figure 2. Shared design principles of both interventions and additional elements of the structure-explicating intervention (Adapted from Peters & Prediger, 2026)

In treating structure explication as a separate additional condition, we refer to Selling's (2016) conceptualization of explicitness in student-centered classrooms, it should not be misinterpreted as direct instruction, but refers to the ways in the ways students are invited and supported to make explicit the underlying ideas, e.g. in teachers' reprising moves that make targeted practices explicit without prescriptively telling them before students start thinking.

When engaging students in rich discourse practices, these were focused on articulating the structures, for which fictitious students gave language models as discursive scaffolds. The lexical scaffolds provided contained structure-explicating phrases, and students were often encouraged to use them, by the tasks as well as teachers' moderation. While these structure-explicating

elements have been identified as supportive in earlier *qualitative* studies (Götze & Baiker, 2021; Peters & Prediger, 2026; Tondorf & Prediger, 2022), no *quantitative* evidence for the hypothesized additional effects on gains in conceptual knowledge has so far been provided.

METHODS OF THE RANDOMIZED CONTROLLED TRIAL

Refined Research Questions and Hypotheses

Summing up the current state of research on arithmetical procedures as presented in the previous section, empirical evidence exists that the design principles of connecting multiple representations and engaging students in rich discourse practices can effectively foster students' conceptual knowledge of arithmetical procedures (Cramer et al., 2002; Götze & Baiker, 2021; Prediger et al., 2022) and determining measurement for area and volume (Battista, 2007). As many interventions show differential efficacy for students with different immigrant and multilingual backgrounds, socioeconomic status, reading proficiency and prior knowledge (Blazar & Archer, 2020; Dietrichson et al., 2017), these background factors need to be taken into account. Procedural knowledge seems to have developed in parallel, and content-related language proficiency for explaining the connections and justifying the procedures seems to be required, so these types of knowledge should also be controlled.

Based on this state of research, we transfer the principles from arithmetical procedures to the procedure of mass unit conversion, and refine the research question to all three types of knowledge and potentially influential background factors. We compared two intervention groups with a control group that worked on a digitally assisted teaching-learning environment in the same platform, yet on another aspect of mass units, namely benchmark knowledge and measurement estimation strategies (Peters et al., 2025; see Section *Independent variable: Interventions with and without structure explication*). We refine the research question to all three types of knowledge and potentially influential background factors:

RQ1. To what extent do fifth graders in the intervention groups (with multiple representations and rich discourse practices) and the control group gain procedural and conceptual knowledge, and content-related language proficiency on mass unit conversion?

RQ2. How are students' learning outcomes related to different treatment conditions?

We formulate the following hypotheses:

H2. Learning gains in mass unit conversion are higher in the intervention groups than in the control group.

Furthermore, we compare the intervention with multiple representations and rich discourse practices to a second intervention that additionally demands and supports the explication of underlying structures, with the third refined research question:

RQ3. Does the structure-explicating intervention have an additional effect on learning gains compared to the regular intervention, and if so, on what exactly?

As structure explication targets the conceptual knowledge and content-related language proficiency but not procedural knowledge, no advantage was assumed for procedural knowledge. Thus, we formulate these two hypotheses for RQ3:

H3a. Learning gains in overall mass unit conversion knowledge are higher in the structure-explicating intervention than in the regular intervention.

H3b. Learning gains in conceptual knowledge and content-related language proficiency are higher in the structure-explicating intervention than in the regular intervention.

Overview on the Research Design

To pursue the research questions under ecologically valid conditions, a classwise randomized controlled trial was conducted with 17 classes, with a total intervention sample of $N = 434$ students who completed the pretest and posttest. As an advance organizer, **Figure 3** presents the pre-post-control group design with three treatment conditions to which classes were randomly assigned, using knowledge for mass unit conversion (split in procedural and conceptual knowledge and content-related language proficiency) as the dependent variable.

Research design of the randomized controlled trial

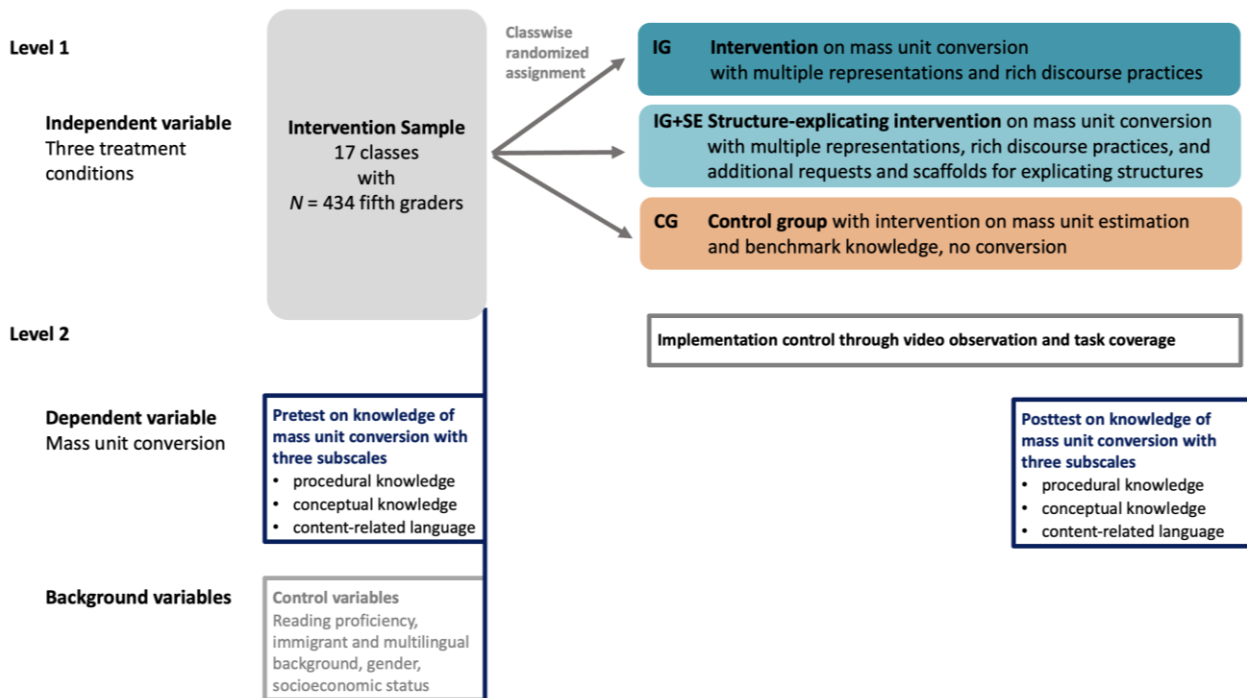


Figure 3. Research design of the classwise randomized controlled trial on mass unit conversion (Source: Authors' own elaboration)

In the following subsections, the methods of data gathering and data analysis are described, following the description of a parallel study on another intervention with which we shared instruments and control groups (Peters et al., 2025).

Independent Variable: Interventions With and Without Structure Explication

Three treatment conditions (one control group and two intervention groups) were compared as dependent variables: Classes in all three conditions worked in a digitally assisted teaching-learning environment on mass units (with 22-32 students each), but only the classes in the intervention groups worked on mass unit conversion.

The *control group* **CG** worked on a digitally assisted teaching-learning environment in the same platform, yet on another aspect of mass units, namely benchmark knowledge and measurement estimation strategies (Peters et al., 2025). By this treatment condition, the motivational effects of working with digital environments were kept constant and a familiarity with mass units, the visual model of the balance scale and the discourse practice of explaining (in this case, estimation strategies) was also achieved, yet without any reference to converting mass units. When discursive goals were comparable, the control group served as baseline condition for conceptual and lexical learning goals.

Both *intervention groups* worked on digital learning environments along the learning trajectory on mass unit conversion as detailed in **Figure 1**, and both followed the design principles to connect multiple representations and engage students in rich discourse practices for three sessions of 45 minutes each. **Figure 2** exemplifies the differences between

- the intervention group **IG** with multiple representations and engaging students in rich discourse practices, and
- the structure-explicating intervention group **IG+SE** with additional requests and scaffolds to explicitly articulate place-value structures, multiplicative bundles structures and quantified refinement structures when connecting multiple representations and justifying the conversion procedure.

In all three conditions, individual and small group work was combined with whole-class discussion in which teachers consistently engaged students in reporting procedures, explaining meanings, describing the visualizations and justifying the procedure and supported their discourse practices by oral discursive and lexical scaffolds. But only in the **IG+SE**, students received support in describing the underlying structures when explaining connections for justifying procedures.

The classes of the intervention and control groups were taught by carefully prepared research-near teachers (highly committed teachers in close cooperation with the university team and daily briefings for each unit). In 13 out of 17 classes, the teachers were new to the classes but assured comparability of the interventions. Teachers of the same treatment conditions received the same teacher preparation (of two hours before the intervention) with intense discussion of the intentions of each task and scaffold and same daily briefings. Teachers of **IG+SE** were trained to engage students in structure-explicating discourse practices and to use the structure-explicating phrases as language models (with concrete moves), request their use from students and leverage their use in the moderation of discussions. Teachers of **IG** and **CG** were trained to request rich explanations and justifications (on unit conversion or on estimation strategies, respectively), but were not allowed to use the structure-explicating phrases which they received only after the posttest.

Treatment fidelity was monitored by video recordings and observation protocols and revealed the complete intended task coverage in all groups in each of the 17 classes. Teachers were found to sufficiently follow the guidelines for classroom interaction, with many prompts to explain in all three treatment groups (**IG**, **CG**, **IG+SE**) and focus on structure-explicating phrases only in **IG+SE**.

Measures for Dependent and Background Variables

Dependent variable: Mass unit conversion pretest and posttest

A test of mass unit conversion with 8 items was conducted before and after the interventions as pretest and posttest to assess the three knowledge types. The items and the scoring scheme are presented in **Figure 4**. The item construction and the scoring scheme is based on the prior qualitative research in the same project (Peters & Prediger, 2026). The validity of the test items was assured in cognitive labs beforehand, working with some students in one-by-one settings engaging them in the same explanations and justifications, the scoring was extensively discussed in expert groups.

- *Procedural knowledge of mass unit conversion* was assessed by four items demanding the symbolic transformation between gram and kilogram with whole numbers or decimals with tenths. A maximum score of 4 could be reached.
- *Conceptual knowledge of mass unit conversion* was assessed by an open item requesting students to explain mass unit conversion, ideally with the blocks on the balance scale, and to justify the procedure. Students' responses were scored according to explicitly addressed structures, the precise/imprecise articulation of representations, and the richness of the discourse practices (example scoring in **Figure 4**). The maximum score was 11.5 for conceptual knowledge. Interrater reliability was tested for 20 % of the data and reached very good reliabilities (Cohen's κ between 0.86 and 0.94).
- *Content-related language proficiency* as assessed by three items that captured how students use, or interpret and explain given structure-explicating language. The scoring depended on the complexity of the demands, with synonyms being also accepted (see **Figure 4**). In total, a maximum score of 4.5 could be reached for content-related language proficiency. Interrater reliability was tested for 20% of the data and reached Cohen's $\kappa = 0.84$ for the open items.

According to the state of research, the three types of knowledge can neither be assumed to be completely independent dimensions nor unidimensionally aligned. We found highly significant but low correlations between procedural knowledge and content-related language proficiency (Pearson's $r = 0.258$ with $p < .001$) and between conceptual and procedural knowledge ($r = 0.134$ with $p < .01$), but not between language and conceptual knowledge ($r = 0.037$ with $p > .05$).

The types of knowledge are weighted unequally, following their relevance in the intervention: conceptual knowledge by a score of 11.5, procedural knowledge by a score of 4, content-related language by 4.5. The total score of 20 will be reported, but the research questions focus on the different types of knowledge.

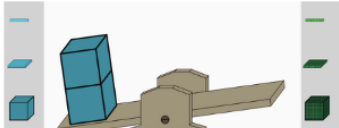
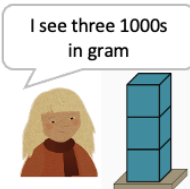
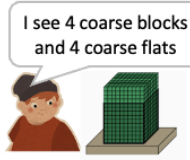
Knowledge domains	Items	Scoring	Scores
Procedural knowledge (Conduct unit conversion)	Four procedural items a) Convert to kilograms: 4000g = _____ b) Convert to kilograms: 3600g = _____ c) Convert to grams: 5kg = _____ d) Convert to grams: 2.4kg = _____	Scoring of each item - Incorrect conversions scored by 0. - Correct conversions scored by 1.	Maximum score of 4
Conceptual knowledge (Explain meaning and justify procedure)	One comprehensive conceptual item Explain how to convert 2kg into grams, for example, using the blocks. Justify why you can do it this way, ideally with the blocks.  Example responses <i>"I believe that 1 kg is 1000 g and 1200 g is 1.2 kg, so 2 kg is 2000. I can do this because 1000 g is 1 kg."</i> <i>"You take a blue block and divide it by 1000 (1000g=1kg). Then you have 1000 small cubes. One small cube equals 1g. You do the same with the second blue block. I can do this because if you cut the blue block into 2000 pieces you actually have 2000g."</i>	Scoring of responses with respect to - Explicitly addressed structures (max. score of 6), - Precise/imprecise articulation of representations (max. score of 4) - Richness of discourse practices (max. score of 1.5) Scoring of example responses - Addressed structures: qualitative refinement structures (1.5) - Representations: symbolic kg & g, no other representation (0.5) - Discourse practice: describing (0.5) - Addressed structures: bundles & quantitative refinement (1.5+3) - Representations: symbolic g & kg; bundle g, visual model g & kg (0.5+1) - Discourse practice: justifying (1.5)	Maximum score of 11.5 <i>Example sum: 2.5</i> <i>Example sum: 9</i>
Content-related language proficiency (Use and explain meaning of phrases)	Three lexical items a) Why does Emma's phrase not match the picture? Explain.  Example response <i>"Because these are coarse blocks and not fine ones. Coarse blocks are kg and the fine blocks are g. So, Emma sees three 1 kg blocks there."</i>	Scoring of responses with respect to - appropriate use and interpretation of structure-explicating language (synonyms were also authorized) - focus on structure-explicating language for bundle structures Scoring of example response - coarse blocks and not fine ones (0.5) - it is kg and not gram (0.25) - there are 1kg-blocks (0.5)	Maximum score of 4.5 <i>Example sum: 1.25</i>
	b) Why does Tom's statement not match the picture? Explain.  Example response <i>"Because they are finely divided blocks and not coarse ones. Same with the flats."</i>	Scoring of responses with respect to - appropriate use and interpretation of structure-explicating language - focus on structure-explicating language for refinement structures Scoring of example response - finely structured not coarse (1) - blocks / flats mentioned (0.25)	<i>Example sum: 1.25</i>
	c) Classify (also sentences and words): What fits to finely structured and what to coarsely structured? coarse c fine c coarse c coarse f fine f 	Scoring of responses with respect to - incorrect assignments scored by 0 - correct assignments scored by 0.125 each Scoring of example response - four falsely assigned (4 x f) (0) - four correctly assigned (4 x c). (0.5)	<i>Example sum: 1</i>
Total score	8 items		Maximum score of 20

Figure 4. All items of the pretest and posttest on mass unit conversion with scoring scheme (Source: Authors' code book)

Table 1. Descriptive characteristics of the intervention sample and the three treatment groups

Variable M(SD) or percent (rounded)	Whole intervention sample (n = 434)	CG Control group with other topics (n = 158)	IG Intervention on mass unit conversion (n = 124)	IG+SE Additionally structure-explicating intervention (n = 152)
<i>Background variables</i>				
Gender (female / male / diverse in %)	43.2 / 48.3 / 0.2	41.1 / 49.1 / 0	45.2 / 45.2 / 0	44.1 / 50 / 0.7
Multilingual background (no / yes in %)	57 / 35.3	50.9 / 40.6	47.6 / 42.7	71.7 / 23
Immigrant background (no / yes in %)	50.8 / 40.4	45.7 / 43.4	45.2 / 45.2	61.2 / 32.9
Socioeconomic status (low / high in %)	49.8 / 41.7	60.5 / 30.2	46 / 44.4	40.8 / 52.6
Reading proficiency (grand-mean centered)	0.12 (4.35)	-1.53 (4.68)	0.22 (4.05)	1.54 (3.71)
Pretest on unit conversion	5.51 (2.97)	4.59 (2.54)	5.62 (3.19)	6.44 (2.93)

Measures of background variables

For capturing potentially relevant background variables, we used the following measures:

- *Reading proficiency.* As mathematical learning gains are often closely linked to reading proficiency (Paetsch et al., 2016), we administered the ELFE II sentence level test, a standardized reading test covering reading speed and comprehension. In the standardization sample, the scale had a high internal consistency (with Cronbach's $\alpha > .90$) and a retest reliability of $r = 0.93$ (Lenhard & Schneider, 2006).

The following background variables were chosen according to their relevance predicting achievement in the German school context in large scale studies (OECD, 2023). They were collected in students' self-report questionnaires preceding the pretest:

- *Gender.* As the German school context continues to produce a small (yet significant) gender gap in mathematics achievement (OECD, 2023, p. 33), the questionnaire captured self-identified gender by asking "Are you a girl/a boy/diverse/no answer".
- *Multilingual background.* Students were asked for the languages spoken at home. Multilingual background was coded for students who reported to speak multiple family languages (or exclusively a non-German family language).
- *Immigrant background.* Students were asked for their and their parents' countries of birth. (Second generation) immigrant background was coded when one of their parents or the students themselves immigrated to Germany.
- *Socioeconomic status.* The family socioeconomic status was captured by the book-at-home index, which has proven to be an economic and reliable self-report instrument ($r = 0.80$; Paulus, 2009). Low socioeconomic status was assigned to all students whose families have max. 100 books and high socioeconomic status for more than 100 books.

Sampling and Samples

The classwise randomized controlled trial was conducted with 17 classes in Grade 5 (students aged 10-12 years, with mean age of 10.40, SD = 0.58). These classes were randomly assigned to the three treatment conditions (see **Figure 3**) with an approximately equal distribution of students. From originally 486 students in the 17 classes, we included those $N = 434$ students in the intervention sample who completed the pretest and the posttest, and whose parents gave written informed consent for the use of anonymized data for research purposes.

Table 1 documents the descriptive characteristics of the whole intervention sample and the three treatment groups. Chi² tests for the background variables revealed significantly fewer students with multilingual background (with $p = .014$) and immigrant background (with $p < .001$) in the intervention groups than in the control group, while gender and socioeconomic status showed no significant difference in pairwise comparisons between the groups (all $p > .05$). Welch-Tests identified higher reading proficiency in the intervention groups (with $p < .001$) and higher results in the pretest scores (Tamhane's-T2 with $p = .09$). In total, the intervention groups started with slightly better prerequisites than the control group. **IG+SE** comprised students with slightly higher average reading proficiency, but had also more students of multilingual and immigrant backgrounds than **IG**. This shows the importance of including the background variables in the later linear models.

Methods of Statistical Analysis

To pursue the descriptive research question **RQ1**, all student answers were rated and scores were descriptively compared between the three treatment groups. For comparing the learning gains for each knowledge type, intra-group effect sizes d were determined by normalizing the mean differences through the pooled standard deviation (pretest across three groups). d counts as small below 0.50, medium between 0.50 and 0.80, large between 0.80 and 1.2, very large above 1.2. Yet, large effects can hardly be expected in short-term interventions in randomized trials (Cheung & Slavin, 2016).

For testing the hypotheses for **RQ2** and **RQ3**, we calculated linear mixed models to account for non-independency of our data. Before conducting the analyses, we considered the possibility of a higher multilevel structure to account for potential class effects from the 17 classes in the sample. As intraclass correlation coefficients (ICC) were very low on the class and the school level and the minimum cluster size for class and schools was the treatment groups were used as the further level, in line with the research question. Therefore, we used a two-level linear mixed model approach, with repeated measurements before and after the treatments, with the treatment group level and individual student. In the linear mixed models, we investigated the effects of the

Table 2. Changes in students' knowledge in mass unit conversion from pretest to posttest – descriptive data for the three treatment groups in three knowledge types and intra-group effect sizes *d* (with pooled *SD*)

	Total test score (max. 20)	Procedural knowledge (max. 4)	Conceptual knowledge (max. 11.5)	Language proficiency (max. 4.5)
<i>Control group CG with other topics</i>				
Pretest <i>M</i> (<i>SD</i>)	4.59 (2.54)	2.43 (1.51)	1.06 (1.62)	1.10 (0.68)
Posttest <i>M</i> (<i>SD</i>)	5.52 (2.91)	2.89 (1.47)	1.34 (1.90)	1.29 (0.66)
Intra-group effect sizes <i>d</i>	0.245	0.351	0.095	0.164
<i>Intervention group IG on unit conversion with multiple representations and rich discourse practices</i>				
Pretest <i>M</i> (<i>SD</i>)	5.62 (3.19)	2.95 (1.52)	1.61 (2.23)	1.05 (0.82)
Posttest <i>M</i> (<i>SD</i>)	9.51 (3.08)	3.59 (1.01)	3.60 (2.25)	2.23 (1.10)
Intra-group effect sizes <i>d</i>	1.003	0.489	0.673	1.139
<i>Intervention group IG+SE with additionally explicating structures</i>				
Pretest <i>M</i> (<i>SD</i>)	6.44 (2.93)	3.24 (1.29)	1.86 (2.36)	1.33 (0.80)
Posttest <i>M</i> (<i>SD</i>)	11.23 (3.47)	3.87 (0.40)	4.51 (3.29)	2.85 (0.72)
Intra-group effect sizes <i>d</i>	1.235	0.367	0.897	1.468

three treatment groups (control group **CG**, intervention groups **IG** or **IG+SE**) on learning gains in mass unit conversion knowledge over time.

To control for individual differences in pretest scores, we set student as random effect resulting in a random intercept for each individual. Individual differences within repeated measurements were accounted for by choosing the repeated measurements design in SPSS. In addition, both interaction terms time x participation in **IG** or **IG+SE** were also included in the model to allow for group-specific learning gains over time. Therefore, we had to include time (0-1 coded) and participation in **IG** or **IG+SE** respectively, dummy-coded with control group **CG** as reference, as additional variables in the models. As fixed effects on the individual level, we entered several time invariant predictors: gender, multilingual background, immigrant background, socioeconomic status and reading proficiency.

Given this model structure, the intercept can easily be interpreted as global mean when all other variables are zero. We tested the assumptions of a linear mixed model – normality of residuals, absence of collinearity, and homoscedasticity – following the guidelines of Meteyard and Davies (2020). Normality was assessed through visual inspection of QQ-plots, while homoscedasticity was evaluated using residual plots. The absence of collinearity was confirmed by calculating the Variance Inflation Factor (VIF). All assumptions were met. We grand-mean-centered values for reading proficiency and dichotomized socioeconomic status by median-splitting before starting the modeling process.

Successive modeling was conducted with SPSS (IBM SPSS Statistics, Version: 28.0.1.0; 142) with ML-estimation starting from a random-effects-only model to more complex mixed models with scaled identity type of covariance for repeated measurements and unstructured covariance for random effects, allowing for assumption-free estimation of covariances and thus highest flexibility. No convergence issues were faced during modeling.

Successive models were compared based on AIC, BIC and LL. P-values were obtained by likelihood-ratio-tests of the full model against the null-model. The final models were then recalculated with REML estimation to get more precise variance estimates (Meteyard & Davies, 2020).

RESULTS

Descriptive Findings on Learning Gains in Three Types of Knowledge

The descriptive data for the research question **RQ1** (To what extent do fifth graders in the intervention groups and the control group gain procedural and conceptual knowledge and content-related language proficiency for mass unit conversion?) is presented in **Table 2** for the total scores and the three knowledge types. The differences between the three treatment groups are first discussed with respect to the difference between control and intervention groups, and then in last subsection with respect to differences between the two intervention groups.

Pretest scores

The pretest scores on *procedural knowledge* were over 50% in all three groups (mean scores 2.43 in **CG**, 2.95 in **IG**, 3.24 in **IG+SE**) because some classes had already learned the procedural skills.

In contrast, the pretest scores on *conceptual knowledge* were very low (mean scores 1.06 to 1.86, with a maximum of 11.5), i.e., the majority of explanations were not rich, rarely connected relevant representations, and rarely addressed relevant underlying structures. Similarly, the low pretest scores for *content-related language proficiency* (mean scores of 1.10 to 1.33, out of max. 4.5) indicate that only very few students could use, interpret and explain structure-explicating phrases. So, the intervention was suitable for classes who had already learned the procedure.

Table 3. Linear mixed models (REML) for predicting students' learning outcomes in posttest with various background variables and treatment groups (compared with control group CG)

Full models	Model T1 <i>Predictors for total posttest score</i>		Model P1 <i>Predictors for procedural test score</i>		Model C1 <i>Predictors for conceptual test score</i>		Model L1 <i>Predictors for language test score</i>	
	Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE
Fixed effects								
Intercept	5.382***	0.339	2.677***	0.146	1.442***	0.259	1.208***	0.086
Time	0.920**	0.288	0.479***	0.109	0.218	0.235	0.205*	0.085
Group level								
Participation in IG+SE	1.397***	0.368	0.708***	0.157	0.547	0.286	0.173	0.096
Participation in IG	0.946*	0.380	0.578***	0.161	0.444	0.295	-0.045	0.099
Participation in IG+SE x time	3.933***	0.411	0.183	0.157	2.515***	0.336	1.287***	0.121
Participation in IG x time	2.955***	0.436	0.148	0.166	1.675***	0.354	0.946***	0.127
Individual level								
Gender (female)	-0.546*	0.253	-0.356***	0.112	-0.201	0.192	0.023	0.062
Immigrant background	0.125	0.311	0.231	0.138	0.117	0.236	-0.216**	0.076
Multilingual background	-0.431	0.324	-0.145	0.143	-0.276	0.246	-0.004	0.080
Low socio-economic status	-0.548*	0.264	-0.297**	0.116	-0.256	0.200	0.027	0.065
Reading proficiency	0.102***	0.030	0.041**	0.013	0.039	0.023	0.026***	0.007
Random effects	Variance	Wald Z	Variance	Wald Z	Variance	Wald Z	Variance	Wald Z
Time	5.719***	13.516	0.838***	13.574	3.881***	13.716	0.498***	13.767
Student (Intercept)	3.217***	6.459	0.772***	8.332	1.558***	5.322	0.116***	3.610
Model fit	-2LL	AIC	-2LL	AIC	-2LL	AIC	-2LL	AIC
	3823	3827	2496	2500	3506	3510	1840	1844
ICC _{conditional}	0.31		0.48		0.29		0.19	

Changes from pretest to posttest scores

In the total scores of the posttest, substantial learning gains were found for the intervention groups **IG** and **IG+SE** (with (very) large intragroup effect sizes of $d = 1.003$ in **IG** and $d = 1.235$ in **IG+SE**), yet a small effect size in the control group **CG** ($d = 0.245$).

The control group **CG** (who worked in a digital teaching-learning environment on estimating units and benchmark knowledge, but not on unit conversion) had small learning effects for procedural knowledge ($d = 0.351$), and content-related language proficiency ($d = 0.164$), yet hardly any in conceptual knowledge ($d = 0.095$). Retest effects might account for a small effect size in procedural knowledge with the same numbers to convert ($d = 0.351$), while the small effect in content-related language proficiency might be traced back to enforced explanations also in the alternative treatment.

In the *intervention groups*, in contrast, the effect sizes differed substantially between the knowledge types. The *procedural knowledge* increased with a slightly higher (yet still small) effect size in **IG** ($d = 0.489$) than in **IG+SE** ($d = 0.367$) and **CG** ($d = 0.351$). The *content-related language proficiency* increased in **IG+SE** with very large effect ($d = 1.468$) and in **IG** with large effect ($d = 1.139$), a strong contrast to **CG** ($d = 0.164$). The conceptual knowledge had a large effect size on **IG+SE** ($d = 0.897$), a medium effect size in **IG** ($d = 0.673$) and none in **CG** ($d = 0.095$). These differences between control group and intervention groups will be shown to be significant in the next sections.

Efficacy of the Interventions Compared to the Control Group

As the control group started with slightly less privileged backgrounds (Table 1), the efficacy of the interventions was analyzed for **RQ2** (How are students' learning gains related to different treatment conditions?) by taking the background factors into account. For this, we tested two hypotheses with four linear mixed models with repeated measurements nested within subjects and students' pre- and posttest scores as dependent variable:

H2. Learning gains in mass unit conversion are higher in the intervention groups than in the control group.

The four models in Table 3 show the estimates derived from the full models for the total test score and the three knowledge types.

Model T1 predicting the total posttest score

For testing **H2 for the total posttest score**, we first report on Model T1. Random effects of time and individual student level accounted for 9% of variance in the data, representing change of scores over time as well as substantial individual differences between students (both with $p < .001$). The main effect of time was highly significant ($p < .01$), meaning that test scores increased significantly from pre- to posttest. The interaction term between time and participation in either one of the intervention groups revealed significant effects (both $p < .001$), that means, in both interventions, students had significantly higher learning gains than in the control group **CG**. The estimates of 2.955 for participation in **IG** x time and 3.933 for participation in **IG+SE** x time can be roughly interpreted as the estimated higher learning gains in the intervention groups than in the control group when controlling for other background variables. So, hypothesis **H2** can be confirmed for the total test score.

Model P1 predicting the procedural posttest score

Procedural knowledge highly significantly increased over time in all three treatment groups (with an estimate of 0.479, $p < .001$) in Model P1, with the same background variables as significant predictors. Yet, adherence to the intervention groups had no

Table 4. Linear mixed model (REML) predicting posttest scores for comparing the structure-explicating intervention **IG+SE** with the regular intervention **IG** (without **CG**)

Full models	Model T2 <i>Predictors for total test score</i>		Model P2 <i>Predictors for procedural test score</i>		Model C2 <i>Predictors for conceptual test score</i>		Model L2 <i>Predictors for language test score</i>	
	Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE
Fixed effects								
Intercept	6.185***	0.403	3.201***	0.144	1.833***	0.337	1.131***	0.108
Time	3.870***	0.364	0.622***	0.123	1.891***	0.302	1.149***	0.108
Group level								
Participation in IG+SE instead of IG	0.374	0.401	0.094	0.142	0.055	0.336	0.221 [†]	0.110
Participation in IG+SE x Time	0.993 [†]	0.489	0.046	0.165	0.843 [†]	0.407	0.339 [†]	0.146
Individual level								
Gender (female)	-0.249	0.318	-0.232 [†]	0.115	-0.059	0.265	0.073	0.082
Immigrant background	-0.252	0.408	0.095	0.147	0.107	0.341	-0.243 [†]	0.106
Multilingual background	-0.212	0.426	-0.113	0.154	-0.096	0.356	-0.027	0.110
Low socio-economic status	-0.419	0.334	-0.219	0.121	-0.257	0.279	0.091	0.086
Reading proficiency	0.168***	0.041	0.065***	0.015	0.083 [†]	0.035	0.021	0.011
Random effects	Variance	Wald Z	Variance	Wald Z	Variance	Wald Z	Variance	Wald Z
Time	7.109***	10.862	0.813***	10.778	5.018***	10.980	0.648***	10.987
Student (Intercept)	2.419***	3.754	0.374***	4.525	1.677***	3.752	0.077	1.629
Model fit	-2LL	AIC	-2LL	AIC	-2LL	AIC	-2LL	AIC
	2464	2468	1462	1466	2319	2323	1259	1262
ICC _{conditional}	0.20		0.32		0.25		0.11	

additional effect on the increases (as the estimates for participation in **IG** and **IG+SE** x time were not significant). A potential ceiling effect with high pretest scores revealed no further significant differences in procedural knowledge gains, and **H2** cannot be confirmed for procedural knowledge.

Model C1 predicting the conceptual posttest score

In Model C1, conceptual knowledge increased highly significantly, but not in the control group ($p > .05$ for time). The estimates for participation x time were highly significant predictors for the two intervention groups (with 2.515 in **IG+SE** x time and 1.675 in **IG** x time, $p < .001$ in both cases). That means that for conceptual knowledge, **H2** can be confirmed.

Meanwhile, none of the background variables were significant predictors in Model C1

Model L1 predicting the language posttest score

In Model L1, content-related language proficiency increased significantly over time in all three groups. The estimates for participation x time were highly significant predictors for the two intervention groups (with 1.287 in **IG+SE** x time and 0.946 in **IG** x time and $p < .001$ in both cases). That means that for content-related language proficiency, **H2** can be confirmed, although not explicitly addressed and supported in **IG**.

For Model L1, reading proficiency and immigrant background were significant predictors, with students whose parents were born outside the country developing less content-related language proficiency.

Additional Effects of the Structure-Explicating Intervention

Both interventions have shown significant efficacy for students' learning gains in mass unit conversion knowledge, so research question **RQ3** asks for the additional effects of the structure-explicating intervention. We compare two intervention groups by testing two hypotheses:

H3a. Learning gains in overall mass unit conversion knowledge are higher in the structure-explicating intervention than in the regular intervention.

H3b. Learning gains in conceptual knowledge and content-related language proficiency are higher in the structure-explicating intervention than in the regular intervention.

The four linear mixed models with repeated measurements in **Table 4** show the estimates derived from the full models for the total test score and the three knowledge types.

Model T2 and P2 predicting the total and procedural posttest score

Model T2 was used to test **H3a** for the **total posttest score**. The main effect of time reconfirmed that test scores in **IG** and **IG+SE** increased significantly from pre- to posttest. The interaction term between time and participation in the structure-explicating intervention (**IG+SE**) revealed significantly higher increases ($p < .05$) than in the regular intervention **IG**, with an estimated difference of 0.993 when controlling for other background variables. So, the hypothesis **H3a** can be confirmed for the total test score.

While **procedural knowledge** highly significantly increased over time in both intervention groups (with an estimate of 0.622, $p < .001$) in Model P2, adherence to the structure-explication intervention group had no additional effect on the increases.

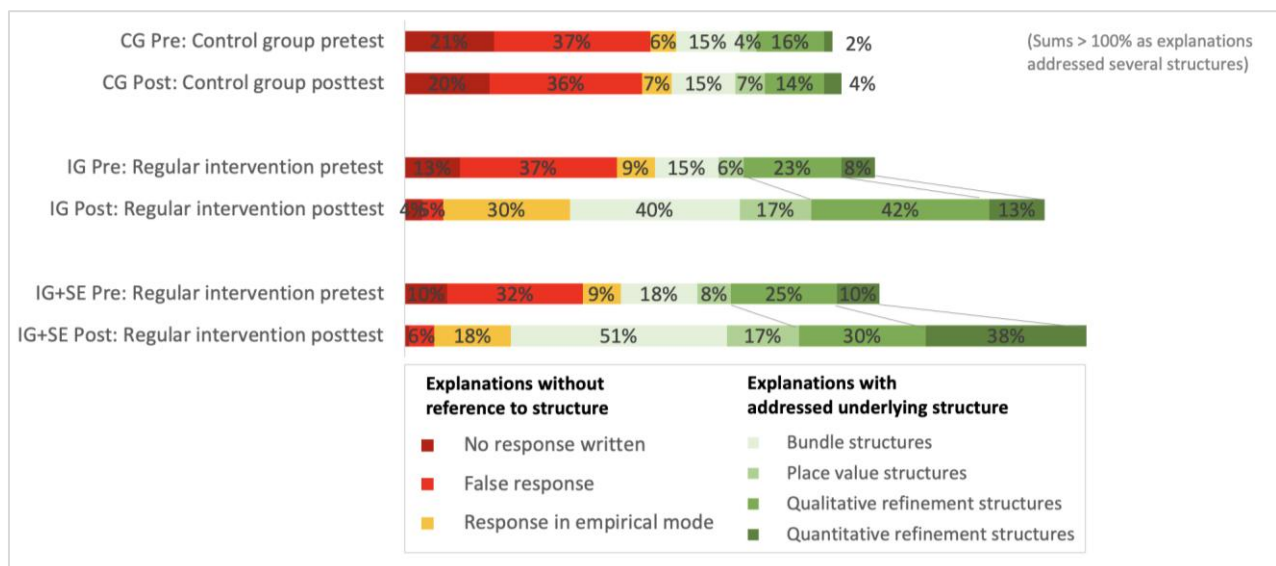


Figure 5. Unpacking conceptual learning gains for three treatment groups: Changed frequencies of addressed structures in students' explanations (Source: Authors' own elaboration)

Model C2 and L2 predicting the conceptual and language posttest score

Model L2 reconfirmed that content-related language proficiency increased significantly over time in both intervention groups (with an estimate of 1.149, $p < .001$). Additionally, the interaction term of participation in **IG+SE** x time was a significant predictor (with 0.339, $p < .05$). So, the hypothesis **H3b** can be confirmed for content-related language proficiency.

Model C2 reconfirmed that conceptual knowledge highly significantly increased over time in both intervention groups (with an estimate of 1.891, $p < .001$). Additionally, the interaction term of participation in **IG+SE** x time was a significant predictor (with 0.843, $p < .05$). So, the hypothesis **H3b** can be confirmed for conceptual knowledge, the additionally structure-explication intervention led to significantly higher conceptual learning gains.

While the additional increases in content-related language proficiency were expected as structure-explicating language was explicitly addressed and supported in the intervention **IG+SE**, the additional effect is less immediate for conceptual knowledge. To better explain why structure-explicating prompts and scaffolds contribute to learning more conceptual knowledge, the following subsection further unpacks the increases in the most relevant aspects of the conceptual knowledge, those related to the addressed structures.

Unpacking the changes in conceptual knowledge of mass unit conversion

The last part of RQ3 asks on *what knowledge aspects exactly* the structure-explicating intervention has an additional effect. Given the results from hypothesis testing, we further unpack the changes in conceptual knowledge as coded in students' explanations for the conceptual item from **Figure 4** in the most relevant subscale for the conceptual knowledge. In **Figure 5**, we compare the changed frequencies in which students of the three treatment groups addressed different structures while explaining why the procedure of converting mass units work in the pretest and the posttest.

While the frequency of explanations without reference to structures remained stable in the control group (21%+37%+6%= 68% in the pretest and 63% in the posttest), it decreased substantially in both intervention groups (from 59% to 39% in **IG** and from 51% to 24% in **IG+SE**).

The structure-explicating intervention group **IG+SE** increased the frequency of addressing *bundle structures* by a factor of 2.8 (from 18% to 51%), while the regular intervention group increased by a factor of 2.7 (from 15% to 40%). The control group changed less than 1%.

In addressing *place value structures*, the **IG** improved slightly more (by factor 2.8, from 6% to 17%) than the **IG+SE** (by factor 2.1, from 8% to 17%), the control group changed less than 1% again.

Refinement structures were identified as most challenging and the most relevant for converting (mass) units in case studies (Peters & Prediger, 2026) and thereby explicated in the design of the structure-explicating intervention in a targeted way, not only in qualitative terms ("this block is finer") but also in quantitative terms ("the 1000-g-block is refiner, it has 1000 cubes, while the 1-kg-block is only one", see **Figure 2**). Whereas the **IG** nearly doubled the addressed *purely qualitative refinement structures* (from 23% to 42%), the **IG+SE** changed much less (from 25% to 30%). In contrast, the **IG** improved only slightly (from 8% to 13%) in *quantified refinement structures* (which might include mentioning qualitative refinement as well), whereas the **IG+SE** increased by factor 3.8 (from 10% to 38%).

These unpacked changes resonate with the design assumption that the quantified refinement structures require the explication more than the other structures, and this led to overall better explanations in the **IG+SE**.

DISCUSSION AND CONCLUSION

Summary and Embedding of Findings

Overall efficacy of fostering mass unit conversion knowledge

In general, the efficacy of the design principles connecting multiple representations (Cramer et al., 2002; NCTM, 2014) and engaging students in rich discourse practices (Erath et al., 2021; Moschkovich, 2015) for fostering conceptual knowledge was already shown for other topics, mainly arithmetical concepts and procedures.

The current controlled trial can replicate empirical evidence of efficacy for the under-researched area of (mass) unit conversion knowledge (Cheeseman et al., 2017), especially for conceptual knowledge of (mass) unit conversion (Tan Sisman & Aksu, 2016; Dincer & Osmanoglu, 2018): The results show that mass unit conversion knowledge can be fostered with large intragroup effect sizes (of $d = 1.235$ in **IG+SE**, $d = 1.003$ in **IG**) for the total test score and with large or medium intragroup effect sizes (of $d = 0.897$ in **IG+SE**, $d = 0.673$ in **IG**) for conceptual knowledge, even in only three lessons of interventions (**Table 2**). As consolidated conceptual knowledge is developed over years by connecting many different aspects, no three-hour intervention can sustainably reach this long-term effect. However, the finding that the short-term interventions can enhance conceptual knowledge scores for mass unit conversion is a promising first contribution to the envisioned long-term development.

When controlling background variables in the linear mixed models, students in both intervention groups showed highly significantly larger learning gains in the total knowledge, in content-related language proficiency and conceptual knowledge of mass unit conversion than the control group who worked also on another digital learning environment on mass units, with a focus on estimating with benchmarks (Peters et al., 2025). In contrast, no significant differences in learning gains were found for procedural knowledge of converting units (**Table 3**).

Beyond the overall evidence of efficacy, the study contributes to theorizing about designs for fostering conceptual knowledge by disentangling the evidence with respect to relevant (a) background variables, and (b) to structure-explication as a key design feature.

Compensating background variables for a more equitable intervention

Different background variables repeatedly occurred in the linear mixed models (**Tables 3** and **4**) as predictive for learning outcomes: gender, reading proficiency and socioeconomic status for *total test score* and *procedural knowledge*, and reading proficiency and immigrant background for *content-related language proficiency*. The effect of reading proficiency is plausible as students were asked to read suggested explanations in the digital teaching-learning environment, and reading counts as a proxy for more general language proficiency which is necessary for explaining mathematical connections (Paetsch et al., 2016). In the future, reading disadvantages might further be compensated in the digital teaching-learning environment, e.g., by reading out loud.

Meanwhile, none of the background variables were substantial predictors for gains in conceptual knowledge, unlike often found in earlier studies (overview in Erath et al., 2021; Paetsch et al., 2016). We interpret this important finding as an indication that our design was able to compensate underprivileged backgrounds by successfully engaging all students in connecting multiple representations (Cramer et al., 2002) and in rich discourse practices (Moschkovich, 2015), with only a very small effect of reading competence left. The discursive and lexical scaffolds can be assumed to have contributed to these compensation effects (Erath et al., 2021) to design a more equitable intervention.

Additional effects of the structure-explicating design features

Explicating structures is an indicator of accomplished conceptual knowledge and presumably an approach for enhancing students' mental acts of structuring and restructuring (Tondorf & Prediger, 2022). Studies on other mathematical concepts have shown an overall relevance of the underlying structures and of articulating them for building conceptual knowledge of concepts and procedures (e.g., Fuson & Briars, 1990; Götze & Baiker, 2021).

But it is only the current study which reduces the lack of empirically disentangling the different principles by explicitly examining the additional effects of supporting students in explicating structures embedded into the design principles of connecting multiple representations and engaging students in rich discourse practices. This intervention study shows that (mass) unit conversion knowledge in total and especially conceptual knowledge can be additionally fostered by eliciting and supporting students in explicating structures: When controlling background variables in the linear mixed models, students in the structure-explicating intervention (**IG+SE**) showed significantly larger learning gains in total unit conversion knowledge and in conceptual knowledge of unit conversion than in the regular intervention (**IG**) without additional structure explication (**Table 4**). We found that students in the intervention **IG+SE** particularly outperformed those in **IG** by addressing more bundle structures and quantitative refinement structures (**Figure 5**), so the fine-grained learning goal specified through the qualitative study (Peters & Prediger, 2026) was reached.

With these results, the controlled trial can contribute new findings on the design principle of explicating structures.

Methodological Limitations and Future Research

While the controlled trial reveals important results, these must be interpreted with caution due to methodological limitations. First, the intervention was taught by relatively inexperienced teachers (although well prepared by the university team) who were not yet very familiar with this specific digitally-assisted approach or with the classes, thus they might have not fully exploited the potentials of the digital teaching-learning environment. In the future, the classroom videos should be further analyzed to examine

different teachers' enactments with respect to teachers' moves and provided oral scaffolds, as earlier studies showed that heterogeneous enactments can reduce the intended differences between interventions (Prediger et al., 2022).

Second, pretest and posttests for the dependent variables were brief, with a ceiling effect for procedural knowledge, while the open item (on explaining why the procedure works) assessed students' conceptual knowledge comprehensively and validly. Future studies should replicate the findings with tests with higher psychometrical reliability and add a retention test for capturing longer-term effects.

Third, classwise random assignment to three treatment conditions did not result in completely comparable student backgrounds, and also pretest scores differed, which was controlled in the linear models. When conducting an additional analysis after propensity score matching to investigate treatment groups of equal sizes and with comparable background factors, the linear models revealed parallel results. However, future studies should examine whether more comparable treatment groups lead to similar results as the unequal prerequisites might have impacted the classroom interaction.

Fourth, slightly unequal group sizes and drop outs in three classes reduced the statistical power of the data. Starting with more classes in the future might allow for better comparability. Trade-offs between individual random assignment and classwise random assignment that assures higher ecological validity due to intact learning groups need to be further negotiated.

Fifth, the short time of the intervention (restricted to three lessons) does not allow for capturing long-term learning gains, so future studies should try to replicate the findings with larger sample sizes and a longer intervention period, especially since building conceptual understanding usually requires consolidation over a longer period of time. Additionally, administering retention tests could assess longer-term learning gains more reliably.

Finally, our findings are still tied to the particular teaching-learning environment, thus their transferability must be investigated in three regards: transfer (a) from converting mass units to other measurement units (such as lengths and areas), (b) from fifth graders to older students, and (c) to many further classrooms to see whether the digital teaching-learning environment also works in scaled-up effectiveness studies under more ecologically valid conditions, with a larger group of teachers less acquainted with the university team (Wortman, 1983).

Implications for Instructional Designs and Classrooms

The study replicates existing empirical evidence that digital teaching-learning environments which engage students in rich discourse practices and connecting multiple representations can substantially foster students' knowledge (Erath et al., 2021), in our case for (mass) unit conversion and especially the conceptual knowledge of this procedure for which mainly qualitative findings existed (Battista, 2007).

The innovative contribution of the study is that the structure-explicating design can contribute to additional learning gains, this is achieved by language models, by eliciting students' explications and supporting students' explanations with structure-explicating scaffolds. The major finding allows us to confirm hypotheses generated in qualitative case studies that enhancing academic language can contribute to meaning making (Erath et al., 2021; Moschkovich, 2015), and allows to focus it much more clearly to a *particular language purpose, the informal yet concise articulation of mathematical structures* (Prediger, 2022). This clear focus can help to tailor other teaching-learning environments to these particular learning needs. In future studies, the principle should be transferred to instructional approaches to foster conceptual knowledge of other procedures.

Author contributions: Spe & SPR: conceptualization, design of teaching-learning environment, writing—original draft; **Spe, SPR, & JW:** methodology, validation, formal analysis, writing—review and editing; **Spe:** data gathering and data curation; **SPR:** supervision, funding acquisition. All authors have read and agreed to the published version of the manuscript.

Funding: The design research project *divomath* is financially supported by the State Ministries of Education of North Rhine-Westphalia and Brandenburg (grant to S. Prediger / C. Selter) and the controlled trial conducted and analyzed for the project Startchancen (BMBF grant 01PL2401C/G to S. Prediger).

Acknowledgements: The authors thank the project partners, especially C. Selter, M. Schröder, and N. Peters, as well as a huge team of student research helpers who supported the data gathering and data analysis process.

Ethical statement: The authors stated that the study was conducted in full accordance with the Declaration of Helsinki of 1975 (on the ethics of the intervention with humans) and in cooperation with the State Department of Education of North Rhine-Westphalia. State authorities (FWU Medieninstitut der Länder and Datenschutz Nord) carefully controlled and approved that the digital teaching learning environment *divomath* fully complies with privacy and data security standards of the General Data Protection Regulation of the European Union. No other data than anonymous data in the digital teaching learning environment were used. Written informed consent was obtained from all teachers and all students' parents in the study.

AI statement: The authors stated that AI was used for language editing.

Declaration of interest: No conflict of interest is declared by the authors.

Data sharing statement: The datasets used and analyzed during the current study are available from the corresponding author on reasonable request.

REFERENCES

- Battista, M. T., & Clements, D. H. (1996). Students' understanding of three-dimensional rectangular arrays of cubes. *Journal for Research in Mathematics Education*, 27(3), 258–292. <https://doi.org/10.2307/749365>
- Battista, M. T. (2007). The development of geometric and spatial thinking. In F. K. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 843–908). Information Age.

- Blazar, D., & Archer, C. (2020). Teaching to support students with diverse academic needs. *Educational Researcher*, 49(5), 297–311. <https://doi.org/10.3102/0013189X20931226>
- Cheeseman, J., McDonough, A., & Golemac, D. (2017). Investigating children's thinking about suspended balances. *New Zealand Journal of Educational Studies*, 52(1), 143–158. <https://doi.org/10.1007/s40841-016-0073-9>
- Cheung, A. C. K., & Slavin, R. E. (2016). How methodological features affect effect sizes in education. *Educational Researcher*, 45(5), 283–292. <https://doi.org/10.3102/0013189X16656615>
- Clark, F. B., & Kamii, C. (1996). Identification of multiplicative thinking in children in grades 1-5. *Journal for Research in Mathematics Education*, 27(1), 41–51. <https://doi.org/10.2307/749196>
- Cramer, K. A., Post, T. R., & delMas, R. C. (2002). Initial fraction learning by fourth- and fifth-grade students: A comparison of the effects of using commercial curricula with the effects of using the rational number project curriculum. *Journal for Research in Mathematics Education*, 33(2), 111–144. <https://www.jstor.org/stable/749646>
- Dietrichson, J., Bøg, M., Filges, T., & Klint Jørgensen, A.-M. (2017). Academic interventions for elementary and middle school students with low socioeconomic status: A systematic review and meta-analysis. *Review of Educational Research*, 87(2), 243–282. <https://doi.org/10.3102/0034654316687036>
- Dincer, E. O., & Osmanoglu, A. (2018). Dealing with metric unit conversion: An examination of prospective science teachers' knowledge of and difficulties with conversion. *Science Education International*, 29(3), 174–182. <https://doi.org/10.33828/sei.v29.i3.5>
- Erath, K., Ingram, J., Moschkovich, J., & Prediger, S. (2021). Designing and enacting instruction that enhances language for mathematics learning – A review of the state of development and research. *ZDM – Mathematics Education*, 53(2), 245–262. <https://doi.org/10.1007/s11858-020-01213-2>
- Freudenthal, H. (1981). Major problems of mathematics education. *Educational Studies in Mathematics*, 12(2), 133–150. <https://doi.org/10.1007/BF00305618>
- Fuson, K. C. (2019). Overcoming errors in fraction computation by emphasizing unit fractions, length drawings, and student explanations. *Universal Journal of Educational Research*, 7(8), 1663–1678. <https://doi.org/10.13189/ujer.2019.070805>
- Fuson, K. C., & Briars, D. J. (1990). Using a base-ten blocks learning/teaching approach for first- and second-grade place-value and multidigit addition and subtraction. *Journal for Research in Mathematics Education*, 21(3), 180–206. <https://doi.org/10.2307/749373>
- Fuson, K. C., Wearne, D., Hiebert, J. C., Murray, H. G., Human, P. G., Olivier, A. I., Carpenter, T. P., & Fennema, E. (1997). Children's conceptual structures for multidigit numbers and methods of multidigit addition and subtraction. *Journal for Research in Mathematics Education*, 28(2), 130–162. <https://doi.org/10.2307/749759>
- Gibbons, P. (2002). *Scaffolding language, scaffolding learning. Teaching second language learners in the mainstream classroom*. Heinemann.
- Glade, M., & Prediger, S. (2017). Students' individual schematization pathways. *Educational Studies in Mathematics*, 94(2), 185–203. <https://doi.org/10.1007/s10649-016-9716-5>
- Götze, D., & Baiker, A. (2021). Language-responsive support for multiplicative thinking as unitizing: Results of an intervention study in the second grade. *ZDM – Mathematics Education*, 53(2), 263–275. <https://doi.org/10.1007/s11858-020-01206-1>
- Hiebert, J., & Lefevre, P. (1986). Conceptual and procedural knowledge in mathematics: An introductory analysis. In J. Hiebert (Ed.), *Conceptual and procedural knowledge: The case of mathematics* (pp. 1–27). Lawrence Erlbaum.
- Hiebert, J., & Wearne, D. (1996). Instruction, understanding, and skill in multidigit addition and subtraction. *Cognition and Instruction*, 14(3), 251–283. https://doi.org/10.1207/s1532690xci1403_1
- Kilpatrick, J., Swafford, J., & Findel, B. (2001). *Adding it up. Helping children learn mathematics*. National Academy Press. <https://doi.org/10.17226/9822>
- Lamon, S. J. (1996). The development of unitizing: Its role in children's partitioning strategies. *Journal for Research in Mathematics Education*, 27(2), 170–193. <https://www.jstor.org/stable/749599>
- Lenhard, W., & Schneider, W. (2006). *ELFE 1-6. Ein Leseverständnistest für Erst- bis Sechstklässler* [ELFE 1-6: A reading comprehension test for first to sixth graders]. Hogrefe.
- Lorange, A., & Rinvold, R. A. (2014). Students' strategies of expanding fractions to a common denominator – a semiotic perspective. *Nordic Studies in Mathematics Education*, 19(2), 57–75.
- Meteyard, L., & Davies, R. A. I. (2020). Best practice guidance for linear mixed-effects models in psychological science. *Journal of Memory and Language*, 112, 104092. <https://doi.org/10.1016/j.jml.2020.104092>
- Moschkovich, J. (2015). Academic literacy in mathematics for English Learners. *The Journal of Mathematical Behavior*, 40(A), 43–62. <https://doi.org/10.1016/j.jmathb.2015.01.005>
- National Council of Teachers of Mathematics [NCTM]. (2014). *Principles to actions. Ensuring mathematical success to all*. NCTM.
- Nguyen, H., An, T., & Maher, E. M. (2023). Impact of designed task-based interventions on k-8 preservice teachers' knowledge and ability to solve measurement unit conversion problems. *International Journal on Studies in Education*, 5(3), 232–244. <https://doi.org/10.46328/ijonse.126>

- Organisation for Economic Co-operation and Development [OECD]. (2023). *PISA 2022 results. The state of learning and equity in education (Volume I)*. OECD. <https://doi.org/10.1787/53f23881-en>
- Paetsch, J., Radmann, S., Felbrich, A., Lehmann, R., & Stanat, P. (2016). Sprachkompetenz als Prädiktor mathematischer Kompetenzentwicklung von Kindern deutscher und nicht-deutscher Familiensprache [Students' growth trajectories in mathematics: The role of language proficiency]. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, 48(1), 27–41. <https://doi.org/10.1026/0049-8637/a000142>
- Paulus, C. (2009). Die "Bücheraufgabe" zur Bestimmung des kulturellen Kapitals bei Grundschulern [The "book-at-home index" for capturing primary students' cultural capital]. <http://hdl.handle.net/20.500.11780/3344>
- Peters, N., Prediger, S., & Weiss, J. (2025). Fostering students' measurement estimation skills in a digital teaching-learning environment: A class-wise randomized controlled trial in Grade 5. *Education Sciences*, 15(2), 238: 1–25. <https://doi.org/10.3390/educsci15020238>
- Peters, S., & Prediger, S. (2026). Enhancing students' understanding of unit conversion by explicating underlying structures: Design research on schematization in Grade 5. *Educational Studies in Mathematics*, 121(2), 283–307. <https://doi.org/10.1007/s10649-025-10437-6>
- Pöhler, B., Prediger, S., & Neugebauer, P. (2017). Content- and language integrated learning: A field experiment for percentages. In B. Kaur, W. K. Ho, T. L. Toh, & B. H. Choy (Eds.), *Proceedings of the 41st Annual Meeting of the International Group for the Psychology of Mathematics Education (PME 41)* (Vol. 4, pp. 73–80). PME.
- Prediger, S. (2022). Enhancing language for developing conceptual understanding: A research journey connecting different research approaches. In J. Hodgen, E. Geraniou, G. Bolondi, & F. Ferretti (Eds.), *Proceedings of Twelfth Congress of the European Society for Research in Mathematics Education (CERME12)* (pp. 8–33). University of Bolzano / ERME. <https://hal.archives-ouvertes.fr/hal-03756062/>
- Prediger, S., Erath, K., Weinert, H., & Quabeck, K. (2022). Only for multilingual students at risk? Cluster-randomized trial on language-responsive instruction. *Journal for Research in Mathematics Education*, 53(4), 255–276. <https://doi.org/10.5951/jresmetheduc-2020-0193>
- Renkl, A., Berthold, K., Große, C. S., & Schwonke, R. (2013). Making better use of multiple representations: How fostering metacognition can help. In R. Azevedo & V. Aleven (Eds.), *International handbook of metacognition and learning technologies* (pp. 397–408). Springer. https://doi.org/10.1007/978-1-4419-5546-3_26
- Rittle-Johnson, B., & Schneider, M. (2015). Developing conceptual and procedural knowledge of mathematics. In R. C. Kadosh & A. Dowker (Eds.), *Oxford Handbook of Numerical Cognition* (pp. 1118–1135). Oxford University Press.
- Ross, S. H. (1989). Parts, wholes and place value: A developmental view. *The Arithmetic Teacher*, 36(6), 47–51.
- Selling, S. K. (2016). Making mathematical practices explicit in urban middle and high school mathematics classrooms. *Journal for Research in Mathematics Education*, 47(5), 505–551. <https://doi.org/10.5951/jresmetheduc.47.5.0505>
- Smith, J. P., & Barrett, J. E. (2017). Learning and teaching measurement: Coordinating quantity and number. In J. Cai (Ed.), *Compendium for Research in Mathematics Education* (pp. 355–385). National Council of Teachers of Mathematics.
- Steinbring, H. (2005). *The construction of new mathematical knowledge in classroom interaction. An epistemological perspective*. Springer. <https://doi.org/10.1007/b104944>
- Tan Sisman, G., & Aksu, M. (2016). A study on sixth grade students' misconceptions and errors in spatial measurement: length, area, and volume. *International Journal of Science and Mathematics Education*, 14(7), 1293–1319. <https://doi.org/10.1007/s10763-015-9642-5>
- Tondorf, A., & Prediger, S. (2022). Connecting characterizations of equivalence. *Educational Studies in Mathematics*, 111(3), 399–422. <https://doi.org/10.1007/s10649-022-10158-0>
- Walshaw, M., & Anthony, G. (2008). The teacher's role in classroom discourse: A review of recent research into mathematics classrooms. *Review of Educational Research*, 78(3), 516–551. <https://doi.org/10.3102/0034654308320292>
- Wessel, L. (2020). Vocabulary in learning processes towards conceptual understanding of equivalent fractions—specifying students' language demands on the basis of lexical trace analyses. *Mathematics Education Research Journal*, 32, 653–681. <https://doi.org/10.1007/s13394-019-00284-z>
- Wortman, P. M. (1983). Evaluation research: A methodological perspective. *Annual Review of Psychology*, 34, 223–260. <https://doi.org/10.1146/annurev.ps.34.020183.001255>
- Zahner, W., Calleros, E. D., & Pelaez, K. (2021). Designing learning environments to promote academic literacy in mathematics in multilingual secondary mathematics classrooms. *ZDM – Mathematics Education*, 53(2), 359–373. <https://doi.org/10.1007/s11858-021-01239-0>