



The role of textbook quality in first graders' ability to solve quantitative comparisons: a multilevel analysis

Henning Sievert¹ · Ann-Katrin van den Ham^{1,2} · Aiso Heinze¹

Accepted: 9 April 2021 / Published online: 18 April 2021
© The Author(s) 2021

Abstract

Students need to create mental models for different types of addition and subtraction situations in order to develop a broad and viable understanding of these operations. Although most students succeed when changing or combining sets, situations that demand a quantitative comparison of sets seem considerably more difficult in the first school year. Textbooks represent the most important learning resource for elementary school mathematics teachers. However, research on their impact on students' knowledge is limited. Hence, we examined textbooks' role in students' ability to model quantitative comparisons by analyzing the learning opportunities presented by four German textbooks for Grade 1 and by conducting a secondary analysis of a dataset based on 1513 students from 84 classes that used one of these textbooks. The results revealed differences in the textbooks' topic-specific instructional quality as well as a significant relation between this quality and student achievement in quantitative comparisons.

Keywords Quantitative comparisons · Textbook quality · Elementary school · Textbook effects

1 Introduction

The main focus of mathematics instruction during the first year of school is ensuring that students obtain a comprehensive understanding of addition and subtraction (Common Core State Standards Initiative, 2010; Kultusministerkonferenz, 2005; Ministry of Education, 2011). Hence, students must learn how to translate different situations that represent an addition or a subtraction problem (semantics) into a mathematical problem (syntax), and also how to assign these situations to a given problem (Kintsch & Greeno, 1985). These addition and subtraction situations are referred to as word problems and are typically classified as change (add to/take away from), combine, equalize, and compare situations (e.g., Van de Walle et al., 2016). However, empirical research suggests that usually only the first three types of problems are dealt with in class or textbooks. In contrast, comparison problems are often neglected or even ignored

(Despina & Harikleia, 2014; Selter et al., 2012; Tarim, 2017). Comparison problems are also considerably more difficult for students in the beginning grades (Stern, 1993). This difficulty might stem from a lack of suitable learning opportunities.

The textbook constitutes the most frequently used learning resource in elementary school mathematics class (Mullis et al., 2012). It frames the teaching activities and learning opportunities provided and is thus considered as potentially implemented curriculum (Schmidt et al., 1997). Previous studies provided first evidence for relations between mathematics textbooks and students' topic-specific achievement (e.g., Schmidt et al., 2001; Sievert et al., 2019, 2021; Törnroos, 2005). Regarding comparison problems, they indicated that the textbook might affect student achievement (Despina & Harikleia, 2014; Tarim, 2017; Xin, 2007). However, the evidence is based on studies with cross-sectional designs and small sample sizes. In view of the relevance of the ability to solve comparison problems for further mathematics learning (Selter et al., 2012), a possible influence of the textbook on the learning success of the students would underline the importance of the role of the textbook as an instrument of change in arithmetic lessons.

Reviewing the state of textbook research, Blazar et al (2019) and Fan (2013) concluded that there is a dearth of

✉ Henning Sievert
sievert@leibniz-ipn.de

¹ Leibniz Institute for Science and Mathematics Education (IPN), Kiel, Germany

² Faculty of Education, Universität Hamburg, Hamburg, Germany

relational and causal studies in this field, and a need for further methodological development. Accordingly, in this study, we examined the relation between textbooks and students' ability to solve comparison problems in Grade 1. Our study is based on a secondary analysis of a dataset comprising relevant data from 1,513 students in 84 school classes that were surveyed during the first seven months of Grade 1. We conducted a theory-based analysis of the four textbooks used in this sample to determine their quality regarding comparison problems. Subsequently, we carried out a multilevel analysis to investigate the relation of the resulting textbook quality to the students' ability to solve comparison problems, while controlling for relevant covariates on the individual and class level.

1.1 Quantitative comparisons

To develop a comprehensive understanding of the addition and subtraction of whole numbers, students must learn to make sense of different addition and subtraction situations, to translate them into a mathematical problem (the mathematical model), and, conversely, to assign them to a given problem (Kintsch & Greeno, 1985). As a translator between these two problems—(real world) word problem and mathematical problem—a student must hence create mental models, as internalized ideas of an action, for different kinds of problems or situations (Riley et al., 1983). Several classifications of such word problems exist, with different subcategories (e.g., Carpenter et al., 2015; Verschaffel et al., 2007; Van de Walle et al., 2016), but all of them distinguish between the situations *change* (add to, take away from), *combine* (part-part-whole), *equalize* (take away from/add to difference), and *compare* (difference). While change and equalize situations are characterized as dynamic (i.e., an action is inherent in the situation), the combine and compare situations are described as static, without any proposals for action (Riley et al., 1983). Further, the quantity in the situation or mathematical model that is unknown can be differentiated (Nunes et al., 2016).

From the four types of situations, comparisons pose the greatest difficulties for students across countries (Nunes et al., 2016; Riley & Greeno, 1988; Stern, 1993). A typical example is as follows: '*Paul has 3 marbles, Mila has 5 marbles. How many more marbles than Paul does Mila have?*' We refer to these problems as quantitative comparisons (QC¹) to distinguish them from qualitative comparisons ('*Paul has 3 marbles, Mila has 5 marbles. Who has more?*'), because the required abilities are distinguishable (Obersteiner et al., 2013). QC are more difficult than other types

of problems, although they can be modeled by an equivalent mathematical model. This difficulty may be due to the static character of a comparison: If the problem itself does not imply an action, to solve it, the students must either generate one themselves, to create a mental model for the problem, or activate a suitable model (such as equalizing). However, the static character alone does not explain the difficulty because the second static situation type, *combine*, is considerably easier for students. Another reason given for the difficulty of QC is the relational nature of the difference in these problems: In contrast to the two sets compared above, the solution to the example task (*2 marbles*) does not describe a concrete, conceivable subset of the given sets but, instead, a relation between two other sets. Stern (1998) described the ability to understand numbers as the ratio between two sets, as a concept that students must develop over time. A third reason for the difficulty that particularly applies to Grade 1 could be problems in text comprehension if the problem is presented in written form (Gabler & Ufer, 2020; Stern, 1993). An example is signal word strategies where the operation is derived directly from the wording of the problem ('less' = subtraction, 'more' = addition, Stern, 1993). Stern (1993) concluded that a lack of understanding of the linguistic symmetry of QC (A has two more than B \Leftrightarrow B has two less than A) and, thus, a missing flexibility in dealing with QC, are possible influencing factors concerning text comprehension in addition to signal word strategies. One more reason, inevitably connected to the previous, is a lack of learning opportunities for comparison problems in class. Several textbook analyses revealed an underrepresentation of QC—or even no representation—in Grade 1 textbooks and, thus, an overrepresentation of the other problem types (e.g., Despina & Harikleia, 2014; Tarim, 2017; Wessel, 2015).

In previous studies, researchers repeatedly reported the long-term value of early numerical competence (e.g., Reeve et al., 2012; Selter et al., 2012) and thus highlighted the need for all aspects of operations, including QC, to be dealt with from the beginning of schooling. However, although the difficulties are known and have been studied for more than three decades, the question of what makes learning environments effective still remains unsatisfactorily answered (Huang et al., 2019). Nonetheless, some exploratory studies gave insights into how to address this issue. Stern (1993) attached importance to linguistic flexibility, particularly students' understanding of the equivalence of 'more than/less than' statements. Similarly, Schumacher and Fuchs (2012) were able to show that students from an intervention focusing on word problems outperformed students from two other groups that focused on calculation and were taught QC conventionally. This intervention effect was partially mediated by the students' understanding of relational terminology. Furthermore, Mwangi and Sweller (1998) reported that studying worked examples was conducive to students

¹ For ease of reading we use the abbreviation QC for both singular and plural.

solving QC, particularly if these examples integrated different representations. Múñez and colleagues (2013) also reported the advantages of additional representations for the construction of mental models for QC. Similarly, the study of Huang et al. (2019) indicated that lessons that were based on a combination of an assumed learning trajectory and variation pedagogy (which focused on using multiple representations amongst other things) had high value for students' ability to solve QC. Múñez et al. (2013) studied students at secondary schools; the other studies considered students from Grades 1–3.

Solving a QC requires several steps, which we outline in an idealized form for the example 'Paul has 3 marbles, Mila has 5 marbles. How many more marbles than Paul does Mila have?' (e.g., Kintsch & Greeno, 1985): First, students must build a situation model, that is, they need to extract the problem and its context from the text: *Two people named Paul and Mila both own a certain number of marbles, Paul 3 and Mila 5. The difference is unknown and needs to be found out.* Second, on this basis, they need to create a mathematical model by abstracting from the concrete persons and objects (marbles) and considering the quantities 3 and 5. The part-whole relation between these quantities, that is, the subset relationship between the two sets represented, has to be identified and established as does the meaning of the difference for solving the problem. The difference and its cardinality can be modeled directly by using manipulatives or by a corresponding addition or subtraction problem, which can be solved by counting or based on number facts (Carpenter et al., 2015). The solution must be interpreted in the situation model. Although this description comprises small steps and some parts of it probably happen implicitly with most students, it gives first insights into the abilities that students need to acquire in order to solve QC.

1.2 Textbooks as learning resources

The textbook is one of the most relevant learning resources for teachers and students in mathematics class. It is conceptualized as a mediator between the official curriculum and the curriculum implemented by teachers because it translates the abstract curriculum into concrete operations that teachers and students can carry out (Valverde et al., 2002). In this paper we use the term curriculum to refer to the level of obligatory learning objectives, which is the most concrete and binding. In some countries such as Germany, national education standards serve as a framework for the whole country; they are adapted by the federal states to different state-specific curricula. In contrast, we consider the term textbook as a possible interpretation and specification of a given curriculum. In Germany, if a textbook is used in one federal state, it generally has to follow the statewide curriculum. Specific data on mathematics textbook use in

schools were collected for TIMSS 2011, which revealed that most elementary school teachers use textbooks as a basis for their instruction. In Germany, where the data of our study were collected, this was the case for 86% of teachers (Mullis et al., 2012).² Only a few quantitative empirical studies exist that examined the relation between mathematics textbooks and teachers' instruction (Krammer, 1985; Schmidt et al., 1997, 2001). They reported a relation between textbooks and teaching practice and lesson content; for instance, topics not covered in a textbook are unlikely to be covered in class (Schmidt et al., 1997).

Looking at the effect of textbooks on student performance, a series of studies exist, most of which indicate such an effect. (Agodini et al., 2010; Bhatt & Koedel, 2012; Bhatt et al., 2013; Hadar, 2017; Koedel et al., 2017; Schmidt et al., 2001; Sievert et al., 2019, 2021; Törnroos, 2005; van den Ham & Heinze, 2018). Although most of these studies were based on black box models, which do not allow conclusions to be made about relevant textbook properties, Schmidt et al. (2001) and Törnroos (2005) found correlations in quantitative textbook characteristics in the U.S. and Finnish TIMSS data, for instance, between the relative share of a topic within a textbook and student achievement in that topic (Schmidt et al., 2001). Further, Sievert et al. (2019, 2021) were able to show a relation between textbook quality and students' strategy use in Grades 1 and 3, providing first criteria for an evidence-based textbook design. However, there are converse results as well. Using publicly available and highly aggregated school-level achievement data, Blazar et al. (2019) did not find differences in student achievement between textbook groups. Their sample included the classes of about 1,200 teachers from six U.S. states, all of which adopted the Common Core State Standards. Similarly, van Steenbrugge et al. (2013) analyzed the data of 1,579 students in Belgium (Grade 1–6) and did not find a relation between the mathematics textbook used and student achievement. As the latter studies just compared textbooks in a black box model, and did not consider the quality of textbooks' characteristics, the state of research on textbook effects on student achievement can be characterized as inconsistent.

Some studies indicated a relation between the mathematics textbook used and student performance in QC. Despina and Harikleia (2014) analyzed the distribution of different types of word problems in Greek elementary school textbooks for Grades 1 and 2 as well as the solutions of 80 students from the first and second grades for these problem types. Their results show that QC were underrepresented in the textbooks analyzed and that the students performed

² However, in a more recent study, Blazar and colleagues (2019) showed that teachers' use of textbooks in general may nevertheless imply significant variance in frequency.

better on the problem types overrepresented in the textbooks than on those underrepresented or missing. Similarly, Tarim (2017) studied the distribution of word problems in six Turkish textbooks (Grades 1–3) as well as the performance of 158 Turkish third-graders on such problems. Her results showed an underrepresentation of QC in the textbooks and greater student difficulties with these problems. Likewise, Xin (2007) examined one U.S. and one Chinese textbook series as well as the performance of 111 students from the United States and China with multiplicative comparisons. She identified an unbalanced distribution of word problem types in the U.S. textbook, in contrast to the Chinese one, and a corresponding unbalanced achievement in the different item types for U.S. students. These cross-sectional studies indicate that the textbook might influence student performance with QC problems. We aimed to extend this research and to examine the relation between textbooks' quality with regard to the topic of QC and student performance. We used a multilevel model with a large sample while also controlling for relevant covariates, collected before and after textbook use.

1.3 Present study

We examined the leverage of the mathematics textbook as a learning resource regarding first graders' ability to solve QC. Our analysis extends the current state of research by using a large-scale database to survey 1513 students from 84 classes in the first seven months of schooling. All textbooks used in this dataset follow the same statewide curriculum; thus, relations between textbooks and student performance are not confounded by varying curricula. Building on the approaches of Törnroos (2005) and Sievert et al. (2019, 2021), we analyzed the learning opportunities in the four different textbook series of our sample. On the basis of previous research results (see Sect. 1.1) and discussions with experts from mathematics education and school practice, we conceptualized textbook quality regarding QC. The resulting scale for textbook quality further develops and refines the research of Despina and Harikleia (2014), Tarim (2017), and Xin (2007), on the relation between textbooks and students' ability to solve QC. Hence, we examined the following research hypotheses:

(1) Hypothesis on textbook quality: First-grade textbooks that follow the same curriculum differ in the quality of the learning opportunities they provide with respect to the different aspects of quantitative comparisons.

(2) Hypothesis on the relation between textbook quality and student performance: The quality of the learning opportunities for quantitative comparisons presented in first-grade textbooks is related to students' ability to solve quantitative comparisons with an unknown difference.

2 Method

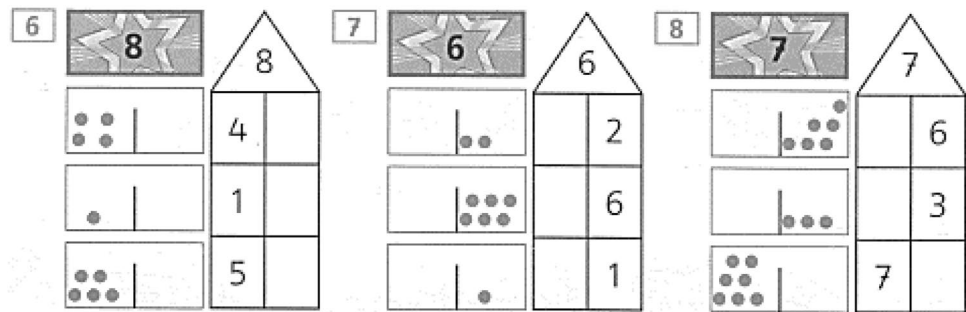
2.1 Participants and design

The study is a secondary analysis of a dataset originally collected for a longitudinal evaluation of a mathematics support program for low-achieving students. It was implemented within regular mathematics classes.³ The evaluation included a comparison of a control group and two intervention groups in which teachers received additional material for low-achievers. This teaching material addressed aspects of basic arithmetic (e.g., concept of cardinal numbers, place-value system, concepts of the basic arithmetic operations) to identify and support these students. It also included material on the part-whole relations of whole numbers in general, and on the inverse relation of addition and subtraction. It did not address QC, and only low-achieving students were supported with the material. The classes that participated in the support program and the textbook that they used were independent of each other. We controlled for possible effects of participation by using dummy-coded variables in the analysis (see Sect. 2.4). It did not show any effect on the student performance in solving QC (see Sect. 3.2).

The original dataset includes the longitudinal data of 2330 students in 127 classes at 40 schools in Schleswig–Holstein (federal state in Germany), throughout elementary school. This is about 10% of the cohort in this federal state. The schools were selected from rural and urban areas all over the federal state. A statewide obligatory curriculum stipulates the content (e.g., number concept, concept and strategies of addition and subtraction), skills (e.g., switch between representation, split numbers), and forms of representations (concrete, iconic, symbolic) to be addressed in the mathematics classroom. Within this dataset, 93 classes used one of the four textbook series “Denken und Rechnen”, “Einstern”, “Flex und Flo”, and “Welt der Zahl”, distributed equally. Other classes used either no textbook or different ones, but this was the case in only a few classes. For this study, nine classes were excluded due to an insufficient number of students ($n < 5$) with data for Grade 1. Further, we excluded 119 students because they did not participate in the QC test. Thus, the sample of this study consists of 1513 students (720 female, 48%; 11 with no data), with a sufficient number of students for each of the four textbook series (Denken und Rechnen, 351; Einstern, 344; Flex und Flo, 485; Welt der Zahl, 333).

³ The studies of van den Ham and Heinze (2018) as well as Sievert et al. (2019, 2021) used data from this dataset as well. The differences in the subsamples of these studies are due to different test instruments being administered at different points of time. The studies focused on different aspects of arithmetic and partly on children in different grades.

Fig. 1 Example for the category of *decomposed numbers with a missing part* (Rinkens et al., 2011); scoring: 3 points (iconic, symbolic, meaningful linking)



The four textbook series are similar in their content structure and follow the same curriculum. The series “Denken und Rechnen” and “Welt der Zahl” include all topics for one school year within one book, while “Flex and Flo” and “Einstein” are divided into three and six books, respectively. All of them cover the same topics for Grade 1 (e.g., beginning with numbers, operations, geometry, patterns). For each textbook, the vast majority of teachers in the sample (90.8–100%) stated that they used the textbook at least once a week for arithmetic instruction.⁴ Furthermore, 81.1–94.1% reported a general orientation towards the content, methods, and chronology of the textbook. Officially, the textbook choice for elementary school in Schleswig–Holstein is made on the school level and not by individual teachers. Little is known about the criteria used (Hartung, 2014). We analyzed whether textbook choice depended on teacher qualification by testing the distributions of textbooks between teachers who had studied mathematics and out-of-field teachers (no such effect was found, see Sect. 3.2).

All data were collected by an authority subordinated to the Ministry of Education in Schleswig–Holstein, assuring the official human subject related approval. We received all raw data fully pseudonymized and we had no contact with students or teachers. The compliance with the human subject related guidelines was monitored by the Data-Protection Supervisor for Schleswig–Holstein.

2.2 Measures

2.2.1 Textbook quality with respect to QC

We developed a measure on the quality of learning opportunities concerning QC presented in the textbooks to analyze textbook quality and to test Research Hypothesis 1. In

⁴ We tested whether the teachers' individual frequency of use was related to textbook quality or student achievement but we did not find pairwise correlations. We conducted further analyses similar to those presented below in Sect. 3.2 and there was neither a main effect of the teachers' specific frequency of use nor an interaction effect of this frequency and textbook quality on student achievement. Hence, we refrained from reporting on these analyses in this paper.

contrast to previous studies (Sievert et al., 2019, 2021), we were unable to build upon an established model, as the question concerning effective QC learning opportunities remains insufficiently answered (Huang et al., 2019). Hence, we deduced first categories from the present state of research and the modeled solution process (see Sect. 1.1), which we discussed with expert teachers from school practice (elementary school mathematics teachers who provide further training for teachers). Their feedback particularly helped us to specify the criteria in each category, for example, in terms of manipulatives and linguistic explanations. Subsequently, we discussed the revised assessment system with researchers in mathematics education. The results of those discussions especially highlighted the need to include the teacher manuals in our analysis, as well as to use the subcategories of introduction, practice, and repetition for the textbook pages. Five categories represent the resulting system for the assessment of the learning opportunities, as follows: decomposed numbers with a missing part, complementarity of addition and subtraction, subtraction as a difference, modeling of QC, and impulses for QC in the teacher manual. We elaborate on the importance of the criteria and their scoring in the following.

2.2.1.1 Category ‘decomposed of numbers with a missing part’ As shown in the modelled solution process (Sect. 1.1), understanding of the relation subset and the associated idea that whole numbers can be decomposed into smaller numbers is crucial for solving QC. In addition, just as in QC, *decomposed numbers with a missing part* usually have a static character (see Fig. 1). We excluded decomposed numbers with a missing whole as they stimulate the mental model for the combine situation.

To assess textbook quality in this category, we first identified textbook pages that presented learning opportunities for *decomposed numbers with a missing part* and divided these into the three subcategories of introduction, practice, and repetition. For each of these subcategories, we then analyzed the learning opportunities and scored one point for each of the following criteria:

- a request for students to use manipulatives;

Fig. 2 Example for the category of *complementarity of addition and subtraction* (Bauer & Maurach, 2011); scoring: 3 points (iconic, symbolic, meaningful linking)

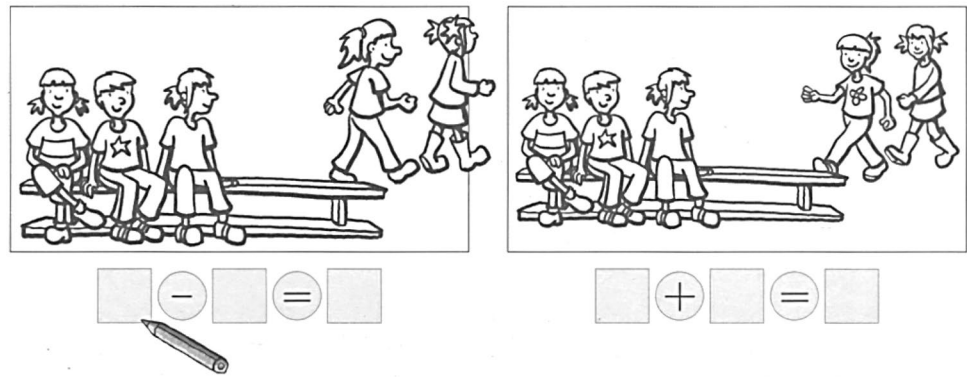
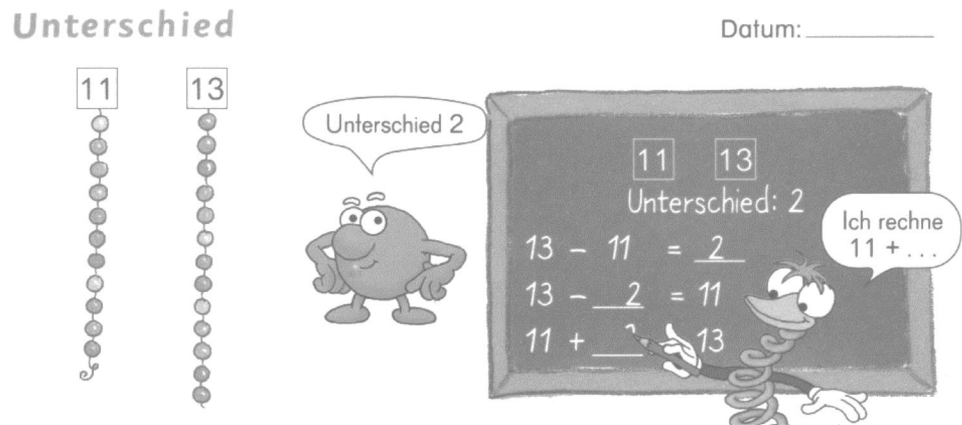


Fig. 3 Example for the category of *subtraction as a difference* (Brall, 2010); scoring: 4 points (iconic, symbolic, linguistic explanation or impulse, meaningful linking)



Note. Translation: Unterschied=Difference. Ich rechne=I calculate.

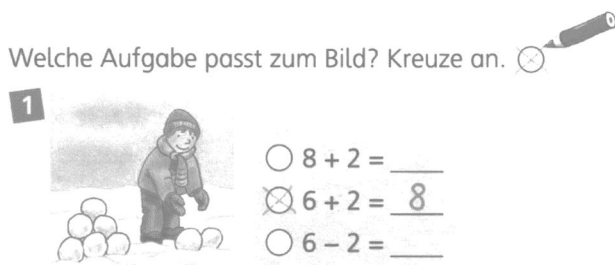
- an iconic representation;
- a symbolic representation;
- a linguistic explanation or impulse (e.g., explain, discuss with others);
- a meaningful linking between these levels of representation.

These criteria take into account former research on the development of strategies to solve QC (Carpenter et al., 2015), on the importance of linguistic flexibility (Schumacher & Fuchs, 2012; Stern, 1993), and on the meaning of multiple representations and their integration (Huang et al., 2019; Muñoz et al., 2013; Mwangi & Sweller, 1998). For each of the subcategories, introduction, practice, and repetition, we assigned 0–5 points. Each textbook could reach 0–15 points in the category *decomposed numbers with a missing part*.

2.2.1.2 Category ‘complementarity of addition and subtraction’ The insight into the *complementarity of addition and subtraction* presents two ways to determine a difference (see Sect. 1.1). It also represents the equivalence of ‘more than/less than’ in a mathematical model, the understanding

of which has been shown to be important for solving QC (Stern, 1993). Examples of corresponding learning opportunities are representations of the invertibility of operations (see Fig. 2) or word problems with two possible interpretations. The scoring was analogous to the previous category (three subcategories, each scored 0–5, possible range of 0–15 points), because the understanding of such relations should be fostered by providing multiple representations as well as their linkage.

2.2.1.3 Category ‘subtraction as a difference’ To understand that the result of a subtraction can represent not only the rest of a take-away-from action but also the difference between the two quantities dealt with, is fundamental for a subtractive approach to QC (Selter et al., 2012). Moreover, the concept of the difference set as the relation between two sets is key in the solution process of QC (see Sect. 1.1). Thus, we rated whether the category of *subtraction as a difference* is addressed in the textbooks. Again, the scoring scheme was the same as for the previous two categories. An example is given in Fig. 3.



Note. Translation: Which task fits the picture? Mark with an x.

Fig. 4 Comparative interpretation as wrong answer (Buschmeier, 2011); scoring: 0 points

2.2.1.4 Category ‘modeling of QC’ Dealing with QC word problems is of course in itself an important learning opportunity for this field (Schumacher & Fuchs, 2012). In textbooks, QC word problems are typically presented in written form or as pictures to be interpreted. We rated whether the textbooks contained such learning opportunities in the category of *modeling of QC*. Because they are presented in one representation that must be translated and put into symbolic form, we did not account for different representations and their linkage in this category. Instead, we used the same three subcategories (introduction, practice, repeat) again, and we scored whether QC word problems were provided in each of them, leading to a possible range of 0–3 points in this category.

2.2.1.5 Category ‘impulses for QC in the teacher manual’ Most learning opportunities in Grade 1 textbooks in Germany are given nonverbally because children are unable to read at this stage. While studying the Grade 1 textbooks, we recognized a series of pictures that could potentially

be interpreted in a QC way, but the context or task usually stimulated other interpretations, such as take-away-from actions. Moreover, Fig. 4 shows a multiple-choice example task, in which the right task should be assigned to an image, while the others should be deemed to be incorrect. The (potential & correct) comparative interpretation here is misrepresented as false. Hence, we decided to examine the teacher manual for possible indications in favor of QC. We did a partial credit scoring in which we decided whether such impulses were not given (0 points), QC impulses were given as a differentiation task for high achievers (1 point), or QC impulses were given for discussion with all students (2 points).

Table 1 gives an overview of the categories, criteria, and scoring of the textbook assessment.

The five category scores were compiled in one textbook quality scale by weighting all categories equally (see Sect. 2.3.1).

2.2.2 Students’ ability to solve QC

We assessed the students’ ability to solve QC with an unknown difference using students’ solutions to QC items from an arithmetic test administered after seven months of schooling. The two item types both included a QC word problem and were identically structured, except that the second type also contained an iconic representation (see Fig. 5).

All items contained static QC problems with an unknown difference. Students were presented with a problem, asked for the difference, and then asked to write down a corresponding problem. The direction of the wording was always positive; how many “more”. The test contained three items with and three without iconic representations, as this could additionally support students.⁵ For each item, we scored 1 point for a correct difference and 2 points for a correct

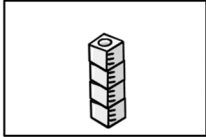
Table 1 Categories, criteria, and scoring of the textbook assessment

Category	Decomposed numbers	Complementarity of add/sub	Subtraction as a difference	Modeling of QC	Teacher manual
Sub-categories	Introduction—practice—repetition				None
Criteria	In each subcategory 1 point each for: use of manipulatives iconic representation symbolic representation request to discuss with others meaningful linking between the representations			In each subcategory: 1 point if QC problems are provided	0 points: no QC impulses given 1 point: QC impulses for high-achievers 2 points: QC impulses for all students
Scoring	0–5 points/ subcategory 0–15 points/ textbook	0–5 points/ subcategory 0–15 points/ textbook	0–5 points/ subcategory 0–15 points/ textbook	0–1 points/ subcategory 0–3 points/ textbook	0–2 points/ textbook

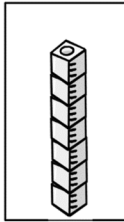
⁵ We found no differences in the scores of items with and without iconic representations.

Fig. 5 Example for the second item type

Tom hat 4 Steckwürfel. Anna hat 7 Steckwürfel.



Tom



Anna

Wie viele Steckwürfel hat Anna mehr als Tom?

Schreibe die Aufgabe dazu auf.

Note. Translation: Tom has 4 cubes. Anna has 7 cubes. How many more cubes than Tom does Anna have? Write down the corresponding problem.

difference in combination with appropriate modeling. In the German mathematics class, the term “Aufgabe” (=problem) in this item context is connoted with a symbolic expression. As appropriate modeling, we accepted each problem that could be used to determine the correct difference, including solutions such as “ $7-4$ ”, “ $4+_ = 7$ ”, “ $7-? = 4$ ”, or “ $4 + 3 = 7$ ”. The actual models given varied a lot across students and classes. In terms of validity, the six items all demanded the translation of and solution to a given QC situation. This required the use and notation of a suitable mathematical model, which the according measure should represent. The range of student solutions, comprising both addition and subtraction modelings, emphasized the different possible solution approaches for these items. The items were tested in a small pilot study and confirmed by an expert elementary school teacher. However, the resulting scale was restricted to the positive wording (‘more’). All appropriate modelings that appeared in the students’ solutions were translated into a syntax and the scoring for both difference and modeling was conducted in a computer-based manner. These scores were totaled for the six items, leading to a scale of 0–12 points, with good reliability (Cronbach’s $\alpha = 0.90$). The scale was standardized and used as a measure for students’ ability to solve QC with unknown differences.

2.2.3 Individual and classroom covariates

We included a series of available individual and classroom-related covariates in our analysis in order to control for possible influencing factors. Various studies have provided evidence that students’ abilities at school entry have an influence on the development of their mathematical skills in elementary school (e.g., Schneider et al., 2013). Hence, we controlled for students’ basic cognitive abilities, basic numerical skills, and German language skills at the beginning of Grade 1. The data were collected by approved standardized tests and yielded adequate reliability measures (basic cognitive abilities: CFT 1-R, Weiß & Osterland, 2013, Cronbach’s $\alpha = 0.91$; basic numerical skills: HaReT 1, Lorenz, 2007, Cronbach’s $\alpha = 0.74$; German language skills: Münsteraner Screening, MÜSC, Mannhaupt, 2013, Cronbach’s $\alpha = 0.72$). The CFT 1-R included 45 items on non-verbal problem solving, the HaReT 1 included 43 items on arithmetic-related precursory skills, and the MÜSC included 40 items testing students’ phonological awareness. The standardized school entry tests were scored as defined in the test manuals. Section 3.1 shows the results for the different textbook series.

On the class level, participation in one of the three groups of the arithmetic support program (see Sect. 2.1)

was included as the dummy variable. We further controlled for class composition effects by including the aggregated variable of basic numerical skills (HaReT 1), because differences among classes could also be caused by different levels of students' average ability. Aggregated basic cognitive abilities and German language skills were excluded due to multicollinearity, which causes distorted estimates. We used a mathematics-specific qualification as an indicator to control for teacher qualification for this covariate (as dummy variable, indicating whether they had studied mathematics or not). Elementary school teachers in Germany are generalists, that is, all teachers teach mathematics but only some have studied mathematics beyond the school level. However, different studies with elementary school teachers revealed that the formal teacher qualification is strongly related to mathematics-specific professional knowledge (e.g., Blömeke et al., 2010; Knievel et al., 2015).

2.3 Procedure

2.3.1 Textbook analysis

Three trained persons independently assessed the four textbooks for Grade 1 based on the first four categories and following the scoring presented in Sect. 2.2.1. Their results indicated substantial agreement (Fleiss's κ , 0.74; ICC, 0.90). Subsequently, we applied the consensus method to score the four textbook categories uniformly. The teacher manuals were reviewed for the fifth category by one trained person. The three QC instructions for the teachers (none, only for high achievers, for all students) were given explicitly in writing, so an interpretation was unnecessary (low-inference rating). To weigh all five categories equally, the resulting scores for each category were standardized and aggregated to a single standardized quality scale on the textbooks' learning opportunities with respect to QC.

2.3.2 Multilevel analysis

Data collection began when the students entered elementary school. Abilities at school entry were assessed at the beginning of Grade 1. The QC items were administered seven months later, in March. Trained test administrators collected the data in accordance with the particular test manuals. All instructions were read aloud. The administrator worked on an example of the QC items with the students. At the end of Grade 1, teachers were given two questionnaires so that data on their qualification and textbook use could be collected.

2.4 Data analysis

To address Research Hypothesis 1, we analyzed differences in textbook quality with respect to QC based on the category

system by comparing and interpreting the scores from the four textbook series.

We applied multilevel random intercept models with the software *Mplus 7.0* (Muthén & Muthén, 1998–2012) to analyze the relation of textbook quality to student performance (Research Hypothesis 2), while taking into account the nested data structure (Hox, 2010). Model 0 without predictors (null model) was run to estimate the partition of variance between and within classes. Model 1 included individual and class composition covariates, namely, students' basic cognitive abilities, basic numerical skills, and German language skills at the individual level, to account for individual characteristics at school entry, and students' basic numerical skills aggregated to a class mean at the group level to control for class composition. Model 2 additionally included the classroom covariates concerning participation in the support program (see Sect. 2.1) and teacher qualification. Model 3 included the textbook quality scale concerning QC. We could not include the five subscales of textbook quality separately due to multicollinearity.

The scores for the individual abilities at school entry, the aggregated class mean of basic numerical skills, and the dependent variable for students' ability to solve QC were standardized, thus the corresponding β -coefficients can be interpreted as effect sizes similar to Cohen's d (Tymms, 2004). A full-information-maximum-likelihood (FIML) approach was applied for missing data on the independent variables. FIML combines missing data and parameter estimation in a single step and uses all the available information (Enders, 2010). Due to the sample selection, there were no missing data for textbooks' quality, the support program, and student performance in QC.

3 Results

3.1 Quality of learning opportunities regarding QC in textbooks

Table 2 presents the results for the textbook quality concerning QC.

None of the four textbooks contained learning opportunities for modeling QC. Although this result was surprising, it is consistent with the results of previous research (e.g., Despina & Harikleia, 2014; Wessel, 2015). One textbook series had the highest scores in all other categories, namely, 'Welt der Zahl', and one series had the lowest score in all but one, namely 'Einstern'. The overall score reflects this result. In contrast, the other two series showed specific strengths and weaknesses. While 'Denken und Rechnen' showed the same score as 'Welt der Zahl' in the *decomposition* category, 'Flex und Flo' had the same score as 'Welt der Zahl' in the *teacher manual* category. Conversely, 'Flex und Flo' had the

Table 2 Results of the textbook analysis: z-scores for the five categories and the overall score (mean)

	Decomposed numbers	Complementarity of add/sub	Subtraction as a difference	Modeling of QC	Teacher manual	Overall score
Denken und Rechnen	0.82	− 0.12	− 0.93	0	− 0.30	− 0.18
Einstern	− 1.63	0.35	− 0.93	0	− 1.51	− 1.28
Flex und Flo	0.00	− 1.50	0.42	0	0.90	− 0.06
Welt der Zahl	0.82	1.27	1.44	0	0.90	1.52

Table 3 Means and standard deviations of the covariates for the different textbook groups (theoretical range below the variable name)

	Denken und Rechnen	Einstern	Flex und Flo	Welt der Zahl
Basic cognitive abilities (0–45)	26.32 (8.62)	25.43 (8.68)	27.54 (9.23)	28.56 (9.04)
Basic numerical skills (0–43)	31.32 (7.01)	31.88 (6.74)	31.90 (7.19)	32.50 (6.57)
German language skills (0–40)	32.77 (6.09)	33.27 (5.92)	33.34 (6.20)	34.46 (5.41)
Basic numerical skills (aggregated) (0–43)	31.29 (1.76)	31.90 (2.41)	31.92 (2.91)	32.52 (3.27)
Support program participation (0–1)	0.82	0.77	0.95	0.63
Teacher qualification (0–1)	0.44	0.43	0.34	0.55

lowest score in the *complementarity* category and ‘Denken und Rechnen’ in the *subtraction as a difference* category. Thus, both manifested a mid-range overall score. Summarizing, one textbook series had high and one had low quality in terms of QC, and two textbooks had medium quality concerning QC.

Compared to the textbook quality determined for arithmetic principles for the same textbooks in a previous study (Sievert et al., 2021), the rankings varied widely. For instance, in representing arithmetic principles, the series ‘Einstern’ showed the highest score. Hence, these results indeed reflect topic-specific and not general textbook quality. Further, a non-topic-specific analysis of textbook effects in arithmetic in another previous study yielded different results as well (van den Ham & Heinze, 2018): students using ‘Denken und Rechnen’ and ‘Welt der Zahl’ outperformed those using ‘Einstern’.

3.2 The relation between textbook quality and students’ ability to solve QC

Table 3 shows the means and standard deviations of all covariates for the four textbook groups. We ran analyses of variance and subsequent post-hoc tests, which showed significant differences for some covariates between the textbook

groups, with small effect sizes of $\eta^2 \leq 0.024$. Hence, it made sense to include these covariates in the multilevel analysis.

The between-class variance in student performance in solving QC as indicated by the null model was 13.7%. Thus, students’ ability to solve QC differed depending on the class the students attended. Table 4 shows the results of Models 1–3 (as described in Sect. 2.4).

All three variables of the students’ abilities had significant main effects on the individual level. They explained about 20% of variance within classes. These effects are in line with recent research on the significance of students’ abilities at the beginning of schooling for their future development of mathematical competence (e.g., Schneider et al., 2013). The results also confirm this finding for QC.

On the class level, the composition according to the aggregated numerical skills did not have an effect. Similarly, participation in the support program did not have an effect, which means that the analysis of textbook correlations was not influenced by the intervention. Also, the teacher qualification did not yield any significant effects.

Regarding Research Hypothesis 2, we found a significant relation between textbook quality and students’ ability to solve QC. For an interpretation of the coefficient ($\beta = 0.09$), we refer to the results of Table 2, which show the range in the textbook quality scale, ranging from − 1.28 to 1.52 (z-scores). Students in classes with the lowest textbook

Table 4 Results of the multilevel analysis of students' ability to solve QC

	Model 1	Model 2	Model 3
Level 1 (students)			
Basic cognitive abilities	0.25** (0.03)	0.25** (0.03)	0.25** (.03)
German language skills	0.15** (0.02)	0.14** (0.02)	0.14** (.02)
Basic numerical skills	0.11** (0.03)	0.11** (0.03)	0.11** (.03)
Level 2 (class)			
Basic numerical skills (aggregated)	0.08 (0.05)	0.08 (0.05)	0.08 (.05)
Support program Group 1		0.05 (0.11)	0.09 (.10)
Support program Group 2		0.14 (0.11)	0.18 (.10)
Teacher qualification		−0.11 (0.08)	−0.13 (0.07)
Textbook quality			0.09** (0.03)
Intercept	0.01 (0.04)	− 0.01 (0.09)	− 0.35 (0.09)
Explained within-class variance	20.3%	20.3%	20.2%
Explained between-class variance	8.3%	13.2%	21.3%

Standard errors are shown in parentheses. All variables are standardized except those for Support program Groups 1 and 2, and for Teacher qualification (0,1), ** $p < 0.01$

quality and students in classes with the highest textbook quality on this scale differed in their ability to solve QC on average by about 0.25 standard deviations. Including the textbook quality led to a considerable increase of explained variance between classes ($\Delta R^2 = 8.1$ percentage points).

4 Discussion

In this analysis, we studied the textbook's role as a learning resource in student achievement in solving QC in Grade 1. For this, we developed a set of indicators of textbook quality for this topic by combining insights from the current state of research and discussions with experts. On the basis of the resulting category system, we analyzed the respective learning opportunities presented in four commonly used mathematics textbooks that follow the same arithmetic curriculum in one German state. We found differences in textbook quality concerning QC. In a subsequent multilevel analysis of student performance in QC with a dataset from 84 classes, comprising the first seven months of schooling, we were able to show a significant and relevant relation between textbook quality and student performance in this field.

4.1 Implications for educational research

Our findings suggest that textbook quality is a relevant predictor for mathematics teaching and learning. Earlier textbook research (see Sect. 1.2) and our own teacher questionnaire data (see Sect. 2.1) suggest that most arithmetic problems in mathematics class stem from or are caused by the textbook used. The results of this study suggest that the quality of textbooks with respect to QC restricts the

corresponding learning environments that teachers implement in their lessons. They also indicate that low textbook quality—reflecting a low range of representations within the categories developed for our assessment—accordingly restricts the variety of addition and subtraction situations addressed in class. This holds true particularly when textbooks and teachers follow the same curriculum. Hence, textbooks should be considered in investigations on teaching effectiveness—even if only one specific mathematical topic is addressed.

Regarding mathematics teachers' education, our results suggest that teachers should recognize quality differences between textbooks, regarding a comprehensive presentation of addition and subtraction, and should compensate for poor quality within a textbook. Thus, they should be able to identify a textbook's weaknesses and to provide students with alternative learning opportunities as part of their pedagogical content knowledge. The categories developed in this study provide tools for this development, which are based on theory as well as on empirical evidence. Consequently, further research on the role of teachers' professional knowledge in textbook choice and use is necessary.

Concerning the understanding of operations, this study highlights the importance of a broad instruction concerning addition and subtraction that uses the whole spectrum of different situation models. Although required by curricula and research (Common Core State Standards Initiative, 2010; Kultusministerkonferenz, 2005; Ministry of Education, 2011; Reeve et al., 2012), QC problems still appear to be underrepresented in classrooms when addition and subtraction are being taught (Despina & Harikleia, 2014; Tarim, 2017; Xin, 2007). Based on data from regular mathematics classrooms, our findings indicate a lack of appropriate learning opportunities in several textbooks for Grade 1, and,

simultaneously, a relation between textbooks and student achievement in solving QC. The results highlight that the lower solution rates of QC problems are not only due to a possible inherent difficulty of the problem type, but also to a lack of learning opportunities in textbooks, limiting those provided in class.

Furthermore, this study contributes to research on the relation between mathematics textbooks and student achievement. The results provide evidence of a relation between textbook choice and student performance in elementary school, in line with the findings of Schmidt et al. (2001) and Törnroos (2005) for secondary school students. We developed specific criteria to assess textbook quality with respect to QC. The development of these criteria was based on empirical research on QC solving and on discussions with experts. The criteria proved to be valid, because based on these we were able to explain an observed relation of textbooks to students' learning outcomes. The results show that a fine-grained, theory-based content analysis of mathematics textbooks is a rewarding approach with which to examine the possible effects of these learning resources. Hence, this study contributes to meeting the need for more relational textbook research, as identified by Fan (2013) or Blazar et al (2019), and extends former research on the significance of topic-specific textbook quality by Sievert et al. (2019, 2021). As this study revealed, it is possible to identify a relation between the learning opportunities presented by textbooks and student achievement in the QC subdomain. This is a useful contribution to research on the quality and impact of mathematics textbooks. It is conceivable that this approach can be transferred to more mathematics topics in order to learn more about the specific strengths and weaknesses of textbooks.

4.2 Implications for educational practice

Our analysis revealed substantial differences in the quality of the learning opportunities presented in four textbooks that follow the same curriculum, and a subsequent relation of these differences to the student performance in the field of QC. Particularly, differences in the range of representations (manipulatives, iconic, symbolic, and linguistic systems, and the relationships among them) for the single category, as well as explicit instructions in the teacher manuals, were shown to lead to differences in textbook quality, which in turn was related to student achievement in solving QC. Hence, this study provides evidence that student achievement not only depends on the curriculum used, but also on the textbook chosen to implement the curriculum. Textbooks are a determinant of student achievement that can be changed easily and cost-effectively. Therefore, they are a valuable instrument for the improvement of student

achievement. They can further be regarded as an instrument of quality assurance in education.

As previously stated, teachers should be aware of quality differences between textbooks regarding QC and should be trained in the assessment of textbook quality. Finally, topic-specific criteria for determining mathematics textbooks' quality, as developed in this study, provide tools for textbook authors and publishing houses with which the strengths and weaknesses of textbooks can be analyzed. These tools could provide valuable opportunities for improving textbook quality.

4.3 Limitations

Some limitations of this study might influence the results and their interpretation. The secondary analysis of an existing dataset that was not based on an experimental setting with respect to textbook distribution provides correlational relations, but does not allow causal interpretations. The sample used in our study, as well as the number of textbooks, could also have had an effect on our results. Further, we were unable to include certain covariates of interest that might possibly interact or be confounded with textbook quality. These include data on the textbook choice in schools, observational data concerning teachers' actual textbook use (instead of teacher reports), or data on teachers' professional knowledge beyond their formal qualifications. All these variables might influence the textbooks' role and should be considered in future studies. The choice of textbooks is particularly crucial for this research field because little is known about the criteria schools or teachers apply. For the United States, this process is described as highly idiosyncratic and market-driven (National Research Council, 2002). For Germany, Hartung (2014) presumed that the process is determined by rash decisions that are influenced by publishers' advertisements or by a close relationship between sales representatives and schools. Hence, the possible influence of specific school characteristics on textbook choice requires additional research. A second limitation is the multicollinearity of the single scales of our category system. The data do not provide details about the weighting of the five categories for QC quality. Although the rankings partly varied among categories, testing all five scales in a single model was not possible, thus we could not draw conclusions about the particular meanings. A third limitation, which is also connected to the secondary analysis, relates to the measure of solving QC. Of course, six items are not a comprehensive basis on which to assess students' ability. The items are restricted to QC problems with an unknown difference set. A more precise analysis of students' performance should also include problems with unknown comparisons or reference sets (cf. Gabler & Ufer,

2020), including their specific difficulties and requirements. However, our measure has solid and reliable psychometric characteristics and the corresponding relation to the textbooks is explainable by the quality of learning opportunities. Nevertheless, the relation found in our study is limited to the field of QC problems with an unknown difference and positive wording. Finally, we analyzed the textbook editions used by the students of our dataset. The results in Table 2 thus might not apply to revised or new editions of the respective textbook series.

5 Conclusion

In this study we reported a significant relation between mathematics textbooks' quality and student performance in solving quantitative comparisons in Grade 1. The results contribute to textbook research by providing theory-driven and topic-specific quality criteria that can be used to assess textbooks. Our findings provide further evidence of the importance of textbook quality for mathematics education in elementary schools. They are based on a large-scale dataset of students from regular mathematics classes, in which the same curriculum was taught. In particular, we were able to show that the topic-specific textbook quality of quantitative comparisons can be determined by a scheme of five categories and that differences are mostly due to the different ranges of representations used within these categories. This topic-specific textbook quality was found to be significantly related to student achievement in solving quantitative comparisons. It thus plays an important role in this context.

Funding Open Access funding enabled and organized by Projekt DEAL.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Agodini, R., Harris, B., Thomas, M., Murphy, R., & Gallagher, L. (2010). Achievement effects of four early elementary school math curricula: Findings for first and second graders. NCEE 2011–4001. National Center for Education Evaluation and Regional Assistance.
- Bauer, R., & Maurach, J. (2011). *Einstern 1*. Cornelsen.
- Bhatt, R., & Koedel, C. (2012). Large-scale evaluations of curricular effectiveness: the case of elementary mathematics in Indiana. *Educational Evaluation and Policy Analysis*, 34(4), 391–412.
- Bhatt, R., Koedel, C., & Lehmann, D. (2013). Is curriculum quality uniform? evidence from Florida. *Economics of Education Review*, 34, 107–121.
- Blazar, D., Heller, B., Kane, T., Polikoff, M., Staiger, D., Carrell, S., & Kurlaender, M. (2019). *Learning by the book: Comparing math achievement growth by textbook in six Common Core states*. Research report. Cambridge, MA: Center for Education Policy Research, Harvard University.
- Blömeke, S., Kaiser, G., & Lehmann, R. (Eds.). (2010). *Professionelle Kompetenz und Lerngelegenheiten angehender Primarstufenlehrkräfte im internationalen Vergleich [TEDS-M 2008- international comparison of future primary school teachers' professional competence and opportunities to learn]*. Waxmann.
- Brall, C. (2010). *Flex und Flo 1*. Diesterweg.
- Buschmeier, G. (2011). *Denken und Rechnen 1*. Westermann.
- Carpenter, T. P., Fennema, E., Franke, M. L., Levi, L., & Empson, S. B. (2015). *Children's mathematics: cognitively guided instruction*. Heinemann.
- Common Core State Standards Initiative. (2010). *Common core state standards for mathematics*. National Governors Association and the Council of Chief State School Officers, Washington, DC. <http://www.corestandards.org/Math/Practice>. Accessed 3 June 2020.
- Despina, D., & Harikleia, L. (2014). Addition and subtraction word problems in Greek grade a and grade b mathematics textbooks: Distribution and children's understanding. *International Journal for Mathematics Teaching and Learning*, 8, 340–354.
- Enders, C. K. (2010). *Applied missing data analysis*. Guilford Publications.
- Fan, L. (2013). Textbook research as scientific research. Towards a common ground on issues and methods of research on mathematics textbooks. *ZDM—Mathematics Education*, 45(5), 765–777. doi:<https://doi.org/10.1007/s11858-013-0530-6>.
- Gabler, L., & Ufer, S. (2020). Flexibilität im Umgang mit mathematischen Situationsstrukturen—Eine Vorstudie für die Entwicklung eines Förderkonzepts zum Lösen additiver Textaufgaben [Flexibility in dealing with mathematical situation structures—a preliminary study for the development of a support concept for solving addition word problems]. *Journal für Mathematik-Didaktik*. <https://doi.org/10.1007/s13138-020-00170-3>
- Hadar, L. L. (2017). Opportunities to learn: Mathematics textbooks and students' achievements. *Studies in Educational Evaluation*, 55, 153–166.
- Hartung, T. (2014). Schulbuchauswahl und Lernmittelfreiheit in den deutschen Bundesländern im Kontext von *Schülerpartizipation* [Textbook choice and the free supply of educational aids in the German federal states in the context of student participation]. Eckert. Working Papers, 11, 1–15. http://www.pedocs.de/volltexte/2015/11061/pdf/EWP_2014_11_Hartung_Schulbuchauswahl.pdf. Accessed 3 June 2020.
- Hox, J. J. (2010). *Multilevel analysis: Techniques and applications. Quantitative methodology series*. Routledge.
- Huang, R., Zhang, Q., Chang, Y., & Kimmins, D. (2019). Developing students' ability to solve word problems through learning trajectory-based and variation task-informed instruction. *ZDM—Mathematics Education*, 51, 169–181.
- Kintsch, W., & Greeno, J. G. (1985). Understanding and solving word arithmetic problems. *Psychological Review*, 92(1), 109–129.
- Kniewel, I., Lindmeier, A. M., & Heinze, A. (2015). Beyond knowledge: Measuring primary teachers' subject-specific competences in and

- for teaching mathematics with items based on video vignettes. *International Journal of Science and Mathematics Education*, 13(2), 309–329. <https://doi.org/10.1007/s10763-014-9608-z>
- Koedel, C., Li, D., Polikoff, M. S., Hardaway, T., & Wrabel, S. L. (2017). Mathematics curriculum effects on student achievement in California. *AERA Open*, 3(1).
- Krammer, H. P. M. (1985). The textbook as classroom context variable. *Teaching and Teacher Education*, 1(4), 273–278. [https://doi.org/10.1016/0742-051X\(85\)90015-0](https://doi.org/10.1016/0742-051X(85)90015-0)
- Kultusministerkonferenz. (2005). *Beschlüsse der Kultusministerkonferenz: Bildungsstandards im Fach Mathematik für den Primarbereich* [Resolutions of the standing conference of the ministers of education and cultural affairs of the federal states of Germany: Educational standards in mathematics for primary education]. https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2004/2004_10_15-Bildungsstandards-Mathe-Primar.pdf. Accessed 3 June 2020.
- Lorenz, J. H. (2007). *HaRet—Hamburger Rechentest für Klasse 1* [Hamburg calculation test for Grade 1]. Freie und Hansestadt Hamburg.
- Mannhaupt, G. (2013). *MÜSC—Münsteraner Screening zur Früherkennung von Lese-Rechtschreibschwierigkeiten* [Münster screening for an early detection of difficulties in reading and writing]. Cornelsen.
- Ministry of Education, P. R. China. (2011). *Mathematics curriculum standards for compulsory education (Grades 1–9) (in Chinese)*. Beijing Normal University Press.
- Mullis, I. V. S., Martin, M. O., Foy, P., & Arora, A. (2012). *TIMSS 2011 international results in mathematics*. TIMSS & PIRLS International Study Center.
- Múñez, D., Orrantia, J., & Rosales, J. (2013). The effect of external representations on compare word problems. Supporting mental model construction. *The Journal of Experimental Education* 81(3), 337–355. doi:<https://doi.org/10.1080/00220973.2012.715095>.
- Muthén, L. K., & Muthén, B. O. (1998–2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.
- Mwangi, W., & Sweller, J. (1998). Learning to solve compare word problems: The effect of example format and generating self-explanations. *Cognition and Instruction*, 16(2), 173–199
- National Research Council. (2002). *Investigating the influence of standards: a framework for research in mathematics, science, and technology education*. The National Academies Press. <https://doi.org/10.17226/10023>
- Nunes, T., Dorneles, B. V., Lin, P.-J., & Rathgeb-Schnierer, E. (2016). *Teaching and learning about whole numbers in primary school*. Springer.
- Obersteiner, A., Reiss, K., & Ufer, S. (2013). How training on exact or approximate mental representations of number can enhance first-grade students' basic number processing and arithmetic skills. *Learning and Instruction*, 23, 125–135. <https://doi.org/10.1016/j.learninstruc.2012.08.004>
- Reeve, R., Reynolds, F., Humberstone, J., & Butterworth, B. (2012). Stability and change in markers of core numerical competencies. *Journal of Experimental Psychology*, 141, 649–666
- Riley, M. S., & Greeno, J. G. (1988). Developmental analysis of understanding language about quantities and of solving problems. *Cognition and Instruction*, 5(1), 49–101
- Riley, M. S., Greeno, J. G., & Heller, J. I. (1983). Development of children's problem-solving ability in arithmetic. In H. P. Ginsburg (Ed.), *The development of mathematical thinking*. (pp. 153–196). Academic Press.
- Rinkens, H.-D., Hönisch, K., & Träger, G. (Eds.). (2011). *Welt der Zahl 1*. Schroedel.
- Schmidt, W. H., McKnight, C. C., Valverde, G. A., Houang, R. T., & Wiley, D. E. (1997). *Many visions, many aims: a cross-national investigation of curricular intentions in school mathematics*. (Vol. 1) Kluwer.
- Schmidt, W. H., McKnight, C. C., Houang, R. T., Wang, H. A., Wiley, D. E., Cogan, L. S., & Wolfe, R. G. (2001). *Why schools matter: a cross-national comparison of curriculum and learning*. Jossey-Bass.
- Schneider, W., Krajewski, K., & Küspert, P. (2013). *Die Entwicklung mathematischer Kompetenzen* [The development of mathematical competence]. Schöningh; UTB GmbH.
- Schumacher, R. F., & Fuchs, L. S. (2012). Does understanding relational terminology mediate effects of intervention on compare word problem. *Journal of Experimental Child Psychology*, 111, 607–628. <https://doi.org/10.1016/j.jecp.2011.12.001>
- Selter, C., Prediger, S., Nührenbörger, M., & Hussmann, S. (2012). Taking away and determining the difference—a longitudinal perspective on two models of subtraction and the inverse relation to addition. *Educational Studies in Mathematics*, 79, 389–408
- Sievert, H., van den Ham, A.-K., Niedermeyer, I., & Heinze, A. (2019). Effects of mathematics textbooks on the development of primary school children's adaptive expertise in arithmetic. *Learning and Individual Differences*, 74(101716), 1–13. <https://doi.org/10.1016/j.lindif.2019.02.006>
- Sievert, H., van den Ham, A.-K., & Heinze, A. (2021). Are first graders' arithmetic skills related to the quality of mathematics textbooks? A study on students' use of arithmetic principles. *Learning and Instruction*, 71(101401), 1–14. <https://doi.org/10.1016/j.learninstruc.2020.101401>
- Stern, E. (1993). What makes certain arithmetic word problems involving the comparison of sets so difficult for children? *Journal of Educational Psychology*, 85, 7–23
- Stern, E. (1998). *Die Entwicklung des mathematischen Verständnisses im Kindesalter* [The development of mathematical understanding in childhood]. Pabst.
- Tarim, K. (2017). Problem solving levels of elementary school students on mathematical word problems and the distribution of these problems in textbooks. *Cukurova University Faculty of Education Journal*, 46(2), 639–648. <https://doi.org/10.14812/cuefd.306025>
- Törnroos, J. (2005). Mathematics textbooks, opportunity to learn and student achievement. *Studies in Educational Evaluation*, 31(4), 315–327
- Tymms, P. (2004). Effect sizes in multilevel models. In I. Schagen & K. Elliot (Eds.), *But what does it mean? The use of effect sizes in educational research*. (pp. 55–66). National Foundation of Educational Research.
- Valverde, G., Bianchi, L. J., Wolfe, R., Schmidt, W. H., & Houang, R. T. (2002). *According to the book: using TIMSS to investigate the translation of policy into practice through the world of textbooks*. Kluwer Academic Publishers.
- Van de Walle, J. A., Karp, K. S., & Bay-Williams, J. M. (2016). *Elementary and middle school mathematics: teaching developmentally*. Person Education Inc.
- Van den Ham, A.-K., & Heinze, A. (2018). Does the textbook matter? Longitudinal effects of textbook choice on primary school students' achievement in mathematics. *Studies in Educational Evaluation*, 59, 133–140. <https://doi.org/10.1016/j.stueduc.2018.07.005>
- Van Steenbrugge, H., Valcke, M., & Desoete, A. (2013). Teachers views of mathematics textbook series in Flanders. Does it (not) matter which mathematics textbook series schools choose? *Journal of Curriculum Studies*, 45(3), 322–353.
- Verschaffel, L., Greer, B., & De Corte, E. (2007). Whole number concepts and operations. In F. K. Lester (Ed.), *Second handbook of research on mathematics teaching and learning*. (pp. 557–628). Information Age.
- Weiß, R. H., & Osterland, J. (2013). *CFT 1-R—Grundintelligenztest Skala 1 Revision* [Basic cognitive abilities test Scale 1 revision]. Hogrefe.

- Wessel, J. (2015): *Grundvorstellungen und Vorgehensweisen bei der Subtraktion* [mental representations and procedures for subtraction]. Dissertation. Fakultät für Mathematik, Technische Universität Dortmund. doi:<https://doi.org/10.1007/978-3-658-11386-5>.
- Xin, Y. P. (2007). Word problem solving tasks in textbooks and their relation to student performance. *The Journal of Educational Research*, 100(6), 347–360. <https://doi.org/10.3200/JOER.100.6.347-360>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.